Hidden Morphophonetic Geometry in Chinese Characters: Sparse Product Lattices, Four-Cycle Order, and Fourth-Corner Prediction

Douglas C. Youvan

doug@youvan.com

www.youvan.ai

June 15, 2026

Chinese characters are often described as combinations of semantic and phonetic components, but this familiar account has rarely been formulated as a global lookup-table geometry. Here we treat the writing system as a sparse product space in which reusable components occupy two functional coordinates: one primarily associated with semantic family and the other with pronunciation. Complete four-character rectangles in the component graph provide self-controlled tests of this organization. When one corner is hidden, the remaining three can often predict the missing character's pronunciation and, more weakly, its semantic field. The effect becomes strongest in the four-cycle-rich core of the graph, suggesting an orderly morphophonetic phase surrounded by a historically irregular lexical periphery. This framework extends earlier work on radicals, phonetic series, character decomposition, and analogical structure while introducing a quantitative order parameter, a fourth-corner completion test, and a direct comparison with the geometry of the genetic code and English morphology. The resulting model does not claim that strokes directly encode meaning. Instead, it proposes that Chinese orthography contains a nonlinear, component-mediated lattice whose topology preserves linguistic information.

Keywords: Chinese characters, Chinese orthography, morphophonetic compounds, semantic components, phonetic components, character lookup table, sparse product lattice, four-cycle topology, fourth-corner prediction, analogical completion, graph theory, nonlinear geometry, component decomposition, phonetic series, semantic radicals, genetic code geometry, English morphology.

OpenAI GPT-5.5 Thinking. CC4.0. 123 pages.

https://orcid.org/0009-0004-4206-4253

Outline

1. Lookup Tables as Objects of Mathematical Discovery

1.1 From Symbol Assignment to Geometry

Introduce the idea that a lookup table may contain more structure than a simple list of input-output assignments. Explain how distances, neighborhoods, symmetries, cycles, and coordinate choices can reveal organization hidden by ordinary tabular presentation.

1.2 The Genetic-Code Precedent

Summarize the motivating genetic-code problem: codons form a complete product space, while amino-acid properties are assigned to the 64 lattice points. Explain how physicochemical and topological organization can be tested against alternative coordinate arrangements.

1.3 The Chinese Character Lookup Table

Define the corresponding Chinese mapping:

structural character description -> character identity, pronunciation, semantic field, and lexical usage.

Emphasize that Chinese differs from the genetic code because the valid character set occupies only a sparse portion of a much larger combinatorial space.

1.4 Central Research Question

State the primary question: does the Chinese character lookup table contain a discoverable geometry in which independent structural coordinates preserve different linguistic outputs?

1.5 Main Hypothesis

Present the working model:

character = Compose(R, P, O, epsilon)

where R is a meaning-associated component, P is a pronunciation-associated component, O is the spatial composition operator, and epsilon is residual historical or graphical variation.

2. Established Chinese Character Structure and the New Mathematical Question

2.1 Semantic and Phonetic Components

Describe the established morphophonetic principle in which one component often indicates a broad semantic class and another often provides a pronunciation cue.

2.2 Radicals, Phonetic Series, and Component Families

Explain the traditional concepts of radicals, phonetic series, recurring components, graphical variants, and historical character formation.

2.3 Why the Known Linguistic Principle Is Not the Present Result

Distinguish the familiar statement that characters have semantic and phonetic components from the stronger mathematical claim that the complete character inventory forms a sparse product lattice containing measurable cycles, axes, and predictive completion laws.

2.4 From Classification to Prediction

Explain why a predictive test is stronger than descriptive classification. A proposed geometry should reconstruct withheld properties or missing cells rather than merely restating known component labels.

2.5 Novelty to Be Evaluated

Identify the potential contribution as the integration of component decomposition, bipartite graph topology, four-cycle density, an order parameter, and held-out fourth-corner prediction.

3. Structural Representation of the Character Inventory

3.1 Character Corpus

Describe the selection of simplified and related Chinese characters, the inclusion criteria, and the treatment of rare, variant, archaic, and incompletely annotated forms.

3.2 Hierarchical Component Decomposition

Represent each character using a composition tree whose leaves are reusable components and whose internal nodes specify spatial operators such as left-right, top-bottom, and enclosure.

3.3 Binary Stroke-Position Representation

Define a lower-level structural code in which each bit indicates the presence or absence of a specified stroke or topological feature within a specified spatial region.

A character is represented by:

$x(c) \in \{0,1\}^D$.

3.4 Multiscale Spatial Features

Explain the use of multiple spatial resolutions to preserve both global arrangement and local distinguishing detail.

3.5 Linguistic Output Tables

Define separate output variables for pronunciation, semantic representation, component identity, radical class, stroke count, historical relation, and lexical usage.

3.6 Separation of Structure and Linguistic Labels

Emphasize that semantic and phonological labels are withheld during construction of the graphical representation so that linguistic organization is tested rather than inserted.

4. The Sparse Morphophonetic Product Lattice

4.1 Two Component Coordinates

For characters with two principal components, define a nominal product space:

R x P,

where R indexes the first structural component and P indexes the second.

4.2 Occupied and Unoccupied Cells

Explain that only a small fraction of theoretically possible component combinations correspond to accepted characters. The observed writing system is therefore a sparse occupied subset of the full product table.

4.3 Bipartite Graph Representation

Construct a bipartite graph whose left vertices are first-coordinate components, whose right vertices are second-coordinate components, and whose edges are existing characters.

4.4 Composition Operators

Separate the graph according to structural operators:

O in {left-right, top-bottom, enclosure, other}.

Discuss why the functional meaning of the first and second coordinates may depend on the operator.

4.5 Connectivity and the Giant Component

Examine whether most productive components belong to one connected network or to many isolated families.

4.6 Productive Core and Sparse Periphery

Introduce the distinction between densely recombined component systems and isolated or historically exceptional characters.

5. Dual Semantic and Phonetic Channels

5.1 First-Component Semantic Affinity

Measure whether characters sharing their first major component have greater semantic similarity than matched controls.

5.2 Second-Component Phonetic Affinity

Measure whether characters sharing their second major component have a greater probability of sharing pronunciation.

5.3 Role Reversal Controls

Compare the expected assignment with the swapped assignment:

first component -> pronunciation
second component -> semantics.

A strong dual-channel structure requires the expected assignment to outperform the reversal.

5.4 Spatial Recovery Without Component Names

Test whether left and right portions of the raw binary stroke-position code recover the same asymmetry without access to conventional component labels.

5.5 The Privileged Decomposition Scale

Compare top-level components, smaller subcomponents, and individual stroke features. Determine whether semantic-phonetic separation reaches a maximum at the complete-component level.

5.6 Operator-Dependent Orientation

Test whether the semantic and phonetic channels rotate or change roles in top-bottom and enclosure structures.

6. Four-Cycle Topology and the Morphophonetic Order Parameter

6.1 Complete Character Rectangles

Define a four-cycle as four occupied cells:

C11 = O(R1,P1)
C12 = O(R1,P2)
C21 = O(R2,P1)
C22 = O(R2,P2).

These four characters form a complete rectangle in component space.

6.2 Four-Cycle Degree

For each character edge, define q as the number of complete rectangles containing that character.

6.3 Morphophonetic Order Parameter

Define an order parameter:

M = semantic affinity through coordinate 1

- semantic affinity through coordinate 2
- phonetic affinity through coordinate 2
- phonetic affinity through coordinate 1.

Positive M indicates the expected semantic-phonetic orientation.

6.4 Transition With Cycle Density

Test whether M increases with q and whether an orderly morphophonetic regime appears above a characteristic level of component recombination.

6.5 Degree-Controlled Null Models

Determine whether cycle density predicts morphophonetic order after controlling for the individual frequencies of both components.

6.6 A Topological Phase Interpretation

Develop the hypothesis that the cycle-rich core constitutes an orderly morphophonetic phase, while the sparse periphery contains reversed, lexicalized, historically opaque, or exceptional constructions.

7. Fourth-Corner Completion

7.1 Definition of the Fourth Corner

Define fourth-corner completion as the prediction of a missing member of a 2 x 2 component rectangle from the other three members.

In the analogy:

A : B :: C : X,

X is the fourth corner or fourth proportional.

7.2 Chinese Semantic Prediction

Use the row-sharing character to predict the missing character's semantic family.

7.3 Chinese Pronunciation Prediction

Use the column-sharing character to predict the missing character's pronunciation.

7.4 Strict Three-Corner Test

Select one supporting rectangle without consulting the hidden character's labels. Use only the three visible corners to generate the prediction.

7.5 Aggregated Local Completion

Combine evidence from all rectangles containing the target character and compare the result with the strict one-rectangle test.

7.6 Swapped-Axis and Randomized Controls

Predict semantics from the phonetic coordinate and pronunciation from the semantic coordinate. Compare these results with label permutations and degree-matched random graphs.

7.7 Representative Character Squares

Present transparent examples in which horizontal movement changes pronunciation while vertical movement changes semantic family.

7.8 Completion Accuracy as Evidence of a Lookup Law

Explain why successful held-out reconstruction provides stronger evidence than clustering, correlation, or visual inspection alone.

8. Nonlinear Geometry of the Character System

8.1 Limits of a Global Linear Map

Explain why a pseudoinverse of the form:

Y_hat = XA

cannot fully represent component interactions, context dependence, alternative layouts, or historical exceptions.

8.2 Conditional Component Functions

Model the output as:

Y = F(R, P, O, epsilon),

where the contribution of one component depends on the identity, position, and operator governing the other component.

8.3 Union of Local Manifolds

Describe the character inventory as a union of locally organized structural regions:

M_total = M_LR union M_TB union M_ENC union M_other.

8.4 Graph Smoothness

Measure whether semantic and phonological labels vary smoothly over the component graph and whether smoothness increases within cycle-rich regions.

8.5 Local Order and Global Irregularity

Explain how strong local prediction can coexist with weak global correlation.

8.6 Historical Residuals

Interpret deviations from the product rule as possible evidence of sound change, semantic drift, graphical simplification, component replacement, borrowing, or lexicalization.

9. Comparison With the Genetic Code

9.1 Complete and Sparse Product Spaces

Contrast the complete 4 x 4 x 4 codon cube with the sparsely occupied Chinese component lattice.

9.2 Input Coordinates and Output Properties

In the genetic code, nucleotide coordinates map to amino-acid identity and physicochemical properties. In Chinese, component coordinates map probabilistically to semantic and phonological properties.

9.3 Degeneracy and Reuse

Compare codon degeneracy with the reuse of semantic and phonetic components across multiple characters.

9.4 Neighborhood Organization

Compare codon adjacency with character rectangles, component sharing, graph neighborhoods, and analogical completion.

9.5 Global Versus Local Order

Explain why the genetic code can exhibit global organization while Chinese orthography may exhibit order primarily in a densely connected local core.

9.6 Limits of the Analogy

Clarify that the genetic code is a biochemical translation system, whereas Chinese orthography is historically evolved, culturally transmitted, polysemous, and subject to language change.

10. Comparison With English Morphological Analogy

10.1 The English Product Table

Represent English inflection as:

lemma x grammatical operation -> word form.

Examples include:

walk : walked :: talk : talked.

10.2 The Fourth Proportional

Relate fourth-corner completion to the established linguistic idea of proportional analogy.

10.3 Productive Rules and Irregular Forms

Contrast highly productive transformations with exceptional forms such as suppletive or irregular past tenses.

10.4 Similarity to the Chinese Lattice

Both systems contain reusable coordinates and permit analogical completion.

10.5 Crucial Difference In English morphology, the axes are explicitly lexical and grammatical. In Chinese, the axes are graphical components whose semantic and phonological functions must be inferred from the lookup-table topology.

10.6 Cross-Linguistic Generalization

Ask whether sparse product lattices and fourth-corner completion provide a general framework for studying writing systems, morphology, and symbolic codes.

11. Implications for Cognition, Artificial Intelligence, and Orthographic Evolution

11.1 Human Learning Through Reusable Components

Discuss how a reader may compress the character inventory by learning recurring semantic and phonetic component families rather than memorizing every glyph independently.

11.2 Analogical Inference

Explain how partial knowledge of a component family could support the pronunciation or interpretation of an unfamiliar character.

11.3 Efficient Artificial Representations

Propose AI representations that preserve component coordinates, composition operators, graph cycles, and historical residuals rather than treating every character as an arbitrary Unicode identifier.

11.4 Explainable Character Embeddings

Show how the product-lattice model could produce interpretable embeddings whose dimensions correspond to component roles and analogical relationships.

11.5 Evolution of Productive Character Families

Consider whether cycle-rich regions reflect historical periods of productive character formation and whether the irregular periphery preserves older or less systematic strata.

11.6 Applications

Discuss possible applications in character education, dictionary organization, optical recognition, pronunciation prediction, rare-character analysis, historical linguistics, and computational paleography.

12. Limitations, Falsification Tests, and Conclusions

12.1 Dependence on Component Decomposition

Discuss how alternative decompositions, graphical variants, and inconsistent component boundaries could alter the inferred lattice.

12.2 Semantic Measurement

Address the limitations of dictionary glosses, translated definitions, polysemy, contextual meaning, and incomplete lexical databases.

12.3 Pronunciation Variation

Consider tones, dialects, historical readings, multiple pronunciations, and sound changes that weaken present-day phonetic correspondence.

12.4 Dataset and Selection Effects

Examine frequency bias, simplified-character bias, missing rare forms, annotation quality, and the treatment of nonproductive components.

12.5 Strong Falsification Tests

Require replication using independent decomposition systems, historical pronunciations, traditional characters, held-out component families, and alternative semantic resources.

12.6 The Strongest Defensible Claim

State that the work does not discover the existence of semantic and phonetic components. It proposes and tests a new global mathematical formulation in which the cycle-rich component graph behaves as a sparse morphophonetic product lattice.

12.7 Final Synthesis

Conclude that the Chinese character lookup table contains an orderly, nonlinear core. In that core, one component coordinate tends to preserve semantic family, the other tends to preserve pronunciation, and complete four-character rectangles act as local decoding circuits capable of fourth-corner prediction.

References

Art

1. Lookup Tables as Objects of Mathematical Discovery

1.1 From Symbol Assignment to Geometry

A lookup table is usually treated as a practical device: one symbol is entered, and a corresponding output is returned. In its simplest form, the table appears to contain no deeper structure. One column lists inputs, another lists outputs, and the relationship between them seems arbitrary. Yet this view can conceal important organization. If the inputs themselves possess internal coordinates, adjacency relations, or compositional rules, then the table can be studied as a geometric object rather than merely as a list.

Suppose the input set is X and the output set is Y. A lookup table is a function

f: X -> Y.

The usual question is: what output f(x) is assigned to a particular input x? A geometric analysis asks additional questions. Which inputs are near one another? Do neighboring inputs tend to have similar outputs? Are there preferred directions in the input space along which particular output properties change? Are there repeated squares, cubes, cycles, symmetries, or low-dimensional surfaces embedded in the table? Do alternative coordinate systems reveal stronger organization than the conventional presentation?

These questions become especially meaningful when X is not an unordered collection but a product space. If

X = X1 x X2 x ... x Xk,

then every input is a coordinate tuple. The table may then be viewed as a field defined over a lattice. The outputs may be numerical values, categories, vectors, pronunciations, meanings, physical properties, or several of these simultaneously.

A lookup table may therefore contain at least three kinds of structure.

First, it may contain local structure: nearby inputs may map to similar outputs.

Second, it may contain global structure: large-scale gradients, clusters, symmetries, or privileged axes may organize the table.

Third, it may contain topological structure: cycles, connected components, bottlenecks, repeated rectangles, and higher-order motifs may constrain how outputs are distributed.

These structures are not necessarily visible in the ordinary ordering of rows. They may emerge only after the inputs are represented in the correct coordinate system.

A central mathematical problem is therefore to distinguish genuine organization from artifacts of representation. Any finite table can be rearranged to appear orderly. A meaningful result requires comparison with null models, alternative coordinate choices, randomized assignments, and held-out prediction. The goal is not to produce an attractive arrangement but to show that one representation reveals statistically unusual and functionally predictive structure.

1.2 The Genetic-Code Precedent

The genetic code provides a particularly clear example of a lookup table with intrinsic coordinates. Each codon contains three nucleotide positions, and each position can take one of four values. The codon space is therefore

{A, C, G, U}^3,

a complete product space containing 4^3 = 64 points.

The lookup table assigns each codon to an amino acid or a stop signal. At first glance, this could be treated as a list of 64 symbolic assignments. However, codons are not arbitrary labels. They occupy a three-coordinate lattice, and codons differing at one position are natural neighbors.

This creates a geometric question: are nearby codons assigned to amino acids with related physicochemical properties more often than expected by chance? One can attach to each codon an output vector describing properties such as hydropathy, volume, charge, or polarity. The genetic code then becomes a field of amino-acid properties defined over codon space.

The coordinate system matters. The three codon positions need not contribute equally, and the four nucleotide values can be ordered in different ways along each axis. Different permutations and anisotropic scalings may reveal different forms of organization. One arrangement may emphasize hydropathy, another may emphasize topology, and another may maximize smoothness under single-nucleotide changes.

The genetic-code problem therefore has several mathematical layers:

codon coordinates, amino-acid assignments, physicochemical output vectors, adjacency relations, graph topology, and alternative lattice orderings.

The important point is methodological. Structure should not be assumed because the table is biologically important. It must be demonstrated by explicit comparison with alternative assignments and coordinate systems. A strong result identifies an arrangement that is unusually smooth, predictive, symmetric, or topologically distinguished relative to a large set of plausible alternatives.

This precedent motivates a broader principle:

A symbolic code may contain hidden geometry when its input symbols are compositional and its output assignments are not random with respect to that composition.

Chinese characters offer a much larger and more complicated test of this principle.

1.3 The Chinese Character Lookup Table

A Chinese character can also be viewed as an input to a lookup table. The output is not a single biochemical object but a lexical record that may include pronunciation, meaning, grammatical function, historical origin, frequency, and relations to other characters.

A simplified representation is

character structure -> linguistic properties.

Unlike codons, Chinese characters are not fixed-length strings drawn from a four-symbol alphabet. They vary in stroke count, component count, spatial arrangement, historical origin, and internal complexity. A character may be composed left-to-right, top-to-bottom, by enclosure, by partial enclosure, or by more irregular arrangements. Components may themselves contain smaller components.

The natural input space is therefore not a complete cube. It is a sparse subset of a much larger combinatorial space.

Let S denote a structural description of a character. This description may include:

stroke types, spatial positions, component identities, composition operators, and topological relations.

The lookup table then maps

S(c) -> L(c),

where L(c) is the linguistic output associated with character c.

The output itself is multidimensional. It can be written schematically as

L(c) = [P(c), M(c), R(c), G(c), H(c)],

where P is pronunciation, M is semantic representation, R is radical or component classification, G is grammatical or lexical usage, and H is historical information.

The space of all possible structural combinations is enormous, but only a small fraction correspond to conventional characters. Thus the Chinese system occupies a sparse set of valid points embedded in a high-dimensional structural space.

This sparsity is not merely an inconvenience. It may be the main source of discoverable organization. Existing characters are not distributed randomly among all possible forms. They were created, transmitted, simplified, borrowed, and modified through historical processes. Productive components were reused many times. Some components became associated with semantic families, while others were reused across pronunciation families.

The relevant object is therefore not just a list of characters. It is a structured lookup table whose occupied points may form clusters, sheets, cycles, rectangles, and product-like subspaces.

1.4 Central Research Question

The central question is:

Does the Chinese character lookup table contain a mathematically discoverable geometry in which different structural coordinates preserve different linguistic properties?

This question is more specific than asking whether visually similar characters are related. It asks whether the entire character system contains reproducible coordinate directions, topological motifs, and local completion laws.

Several subsidiary questions follow.

Do characters sharing one major component tend to belong to related semantic fields?

Do characters sharing another major component tend to have related pronunciations?

Does this separation depend on the spatial composition operator?

Are complete component rectangles unusually common?

Does the number of such rectangles surrounding a character predict how orderly its semantic and phonetic behavior will be?

Can the missing member of a four-character rectangle be predicted from the other three?

Does the structural system contain an orderly core and an irregular periphery?

These questions turn a traditional linguistic observation into a mathematical lookup-table problem.

The known linguistic principle is that many characters are morphophonetic compounds. One component often suggests a semantic domain, while another often suggests pronunciation. The new question is whether this principle produces a measurable product geometry across the full character inventory.

1.5 Main Hypothesis

The working hypothesis is that many Chinese characters can be approximated by

character = Compose(R, P, O, epsilon),

where R is a meaning-associated component, P is a pronunciation-associated component, O is the spatial composition operator, and epsilon represents residual variation.

The residual term includes historical sound change, semantic drift, graphical simplification, irregular component substitution, lexicalization, and exceptions.

For the dominant left-right class, the model often takes the form

character = LR(R, P) + epsilon.

The corresponding lookup-table hypothesis is

meaning family approximately follows R,

pronunciation family approximately follows P.

This does not mean that R fully determines meaning or that P perfectly determines pronunciation. The claim is probabilistic and local. Components provide structured tendencies, not exact outputs.

The hypothesis predicts that characters sharing R should display semantic affinity, while characters sharing P should display phonetic affinity. It also predicts that complete rectangles of the form

C11 = Compose(R1, P1)

C12 = Compose(R1, P2)

C21 = Compose(R2, P1)

C22 = Compose(R2, P2)

should behave like local product cells.

Moving along one direction of the rectangle changes the phonetic coordinate while approximately preserving semantic family. Moving along the other direction changes the semantic coordinate while approximately preserving pronunciation.

The strongest test is not whether such patterns can be found after inspection. It is whether they can predict withheld information. If C22 is hidden, the other three corners should constrain both its likely semantic field and its likely pronunciation.

The deeper hypothesis is that this predictive regularity is concentrated in the cycle-rich core of the component graph. Characters involved in many complete rectangles should behave more systematically than isolated characters.

This leads to the central model of the paper:

The Chinese character system contains a sparse morphophonetic product lattice. Its orderly regions are defined not merely by visual similarity but by repeated component recombination, dual semantic and phonetic channels, and four-cycle topology.

2. Established Chinese Character Structure and the New Mathematical Question

2.1 Semantic and Phonetic Components

A large fraction of Chinese characters are traditionally understood as compounds in which one part contributes information about meaning and another contributes information about pronunciation. The semantic component often places the character within a broad conceptual field, while the phonetic component often suggests how the character was pronounced when the form was created or standardized.

A common structural pattern can be written as

character = semantic component + phonetic component.

This expression is only approximate. The semantic contribution may be broad, indirect, historically altered, or difficult to recognize in modern usage. The phonetic contribution may reflect an older pronunciation rather than the current pronunciation in standard Mandarin. Even so, the recurring division between semantic and phonetic roles is one of the central organizing principles of the writing system.

For many left-right characters, the semantic component appears on the left and the phonetic component appears on the right. This is not a universal law, but it is a frequent pattern. Other characters place the semantic component above, below, outside, or inside the phonetic component. The location depends on the structural operator used to compose the character.

A familiar example is the reuse of a water-related component in characters associated with liquids, rivers, washing, flow, or moisture. The remaining component may contribute a pronunciation pattern. Similarly, characters involving plants, metals, speech, movement, illness, or the body often reuse recurring semantic components associated with those conceptual fields.

The semantic and phonetic functions are not equally reliable. A component may provide a strong pronunciation clue in one character family but a weak clue in another. Semantic components often identify only a broad category rather than the exact meaning of the completed character. Historical sound change, semantic drift, graphical simplification, dialectal variation, and borrowing can all weaken the original relationships.

The established linguistic principle is therefore probabilistic:

P(related meaning | shared semantic component) is greater than chance,

and

P(related pronunciation | shared phonetic component) is greater than chance.

The present study does not begin by disputing this principle. It begins by asking whether the principle produces a measurable geometry over the entire character inventory.

2.2 Radicals, Phonetic Series, and Component Families

Traditional dictionaries often organize characters by radicals. A radical is a selected component used for indexing, classification, or dictionary lookup. In many cases it is also semantically informative, but the dictionary radical is not always identical to the historically meaningful component or the most useful graphical component.

The distinction matters because a character may contain several recognizable parts. One part may be chosen as the dictionary radical even when another part carries stronger semantic information. Conversely, a recurring graphical unit may function as a phonetic component without being treated as the radical.

A phonetic series is a family of characters that share a component associated with pronunciation. Members of the same series may differ in semantic component while retaining related readings. A simplified abstract form is

C1 = Compose(R1, P)

C2 = Compose(R2, P)

C3 = Compose(R3, P),

where P is the shared phonetic component and R1, R2, and R3 are different semantic components.

The corresponding semantic family has the opposite form:

C1 = Compose(R, P1)

C2 = Compose(R, P2)

C3 = Compose(R, P3),

where R is shared while the phonetic components vary.

These two family types suggest that the character inventory can be organized by two partially independent relations:

same semantic coordinate,

same phonetic coordinate.

A character belongs simultaneously to both kinds of family. It occupies the intersection of one semantic-component class and one phonetic-component class.

This intersection structure is more informative than a simple tree. A tree can show that a character contains particular parts, but it does not automatically show repeated recombination across many families. The same semantic component may combine with many phonetic components, and the same phonetic component may combine with many semantic components.

The system is therefore naturally represented as a network or partial product table.

If R is the set of semantic-side components and P is the set of phonetic-side components, then the theoretical product space is

R x P.

Each existing character occupies one cell in this product. Most possible pairs do not correspond to accepted characters, so the table is sparse.

This sparse table contains the material needed for a global mathematical analysis. Rows correspond approximately to semantic families. Columns correspond approximately to phonetic families. Repeated recombination creates rectangles, cycles, and connected regions.

2.3 Why the Known Linguistic Principle Is Not the Present Result

It would not be new to state that Chinese characters often contain meaning-associated and sound-associated components. That observation has been recognized for centuries and is central to traditional classifications of character formation.

The present question is different.

The traditional statement is descriptive:

many characters have a semantic component and a phonetic component.

The mathematical statement proposed here is structural:

the full character inventory forms a sparse product geometry whose two coordinates preserve different linguistic outputs and whose local cycles support prediction.

These are not equivalent claims.

A descriptive account can classify one character at a time. It can say that a particular component is semantic and another is phonetic. It does not necessarily determine whether the entire character system contains a coherent lattice, whether the same directional roles recur across thousands of characters, or whether the strength of those roles depends on graph topology.

The proposed model also introduces a distinction between isolated examples and repeated recombination. A single character can be analyzed as a semantic-phonetic compound, but one character does not define a product space. A product space emerges when components are reused in multiple combinations.

For example, suppose four characters exist:

C11 = Compose(R1, P1)

C12 = Compose(R1, P2)

C21 = Compose(R2, P1)

C22 = Compose(R2, P2).

These four characters form a complete rectangle. The rectangle supplies internal comparisons that are unavailable from a single character.

The two characters in the same row share the first coordinate. The two characters in the same column share the second coordinate. If row-sharing predicts semantic similarity and column-sharing predicts pronunciation similarity, then the rectangle behaves as a local morphophonetic product cell.

The novelty question therefore concerns the full geometry, not the existence of components.

The potentially new contribution is the demonstration that:

complete component rectangles are abundant,

their linguistic roles are directionally asymmetric,

cycle-rich regions behave more regularly than isolated regions,

and withheld character properties can be predicted from neighboring cells.

2.4 From Classification to Prediction

Classification describes what is already known. Prediction tests whether the proposed structure has explanatory power.

Suppose a character has already been labeled with a semantic component and a phonetic component. It is easy to report that the semantic side relates to meaning and the phonetic side relates to sound. That result may simply repeat the original annotation.

A stronger test hides part of the information.

If one character in a complete component rectangle is withheld, the model must infer its properties from the remaining three corners.

Let

C11 = Compose(R1, P1)

C12 = Compose(R1, P2)

C21 = Compose(R2, P1)

C22 = unknown.

The other three corners imply that the missing form should be associated with R2 in one direction and P2 in the other.

The semantic prediction is derived from the row:

C21 and C22 share R2.

The phonetic prediction is derived from the column:

C12 and C22 share P2.

The test is successful only if the hidden character's actual meaning family and pronunciation agree with these predictions more often than appropriate controls.

Prediction reduces the risk of circular reasoning. The hidden output is not used to define the rule applied to that character.

The strongest version of the test uses only one supporting rectangle. A broader version combines information from all rectangles containing the target. Both versions should be compared with swapped-axis controls, randomized labels, degree-matched graphs, and frequency-matched component pairs.

A predictive rule can also fail in informative ways. Failure may indicate an irregular character, historical sound change, semantic drift, mistaken decomposition, or a component whose role is not morphophonetic.

Thus exceptions are not merely noise. They may identify boundaries between orderly and irregular regions of the lookup table.

2.5 Novelty to Be Evaluated

The novelty of this work must be stated carefully.

The project does not claim to discover radicals, phonetic series, semantic components, or morphophonetic compounds. It does not claim that all Chinese characters obey a simple two-part rule. It does not claim that visual structure uniquely determines meaning.

The proposed contribution is a global mathematical formulation with several linked elements.

First, the character inventory is represented as a sparse component product table.

Second, the table is converted into a bipartite graph in which characters are edges connecting two component coordinates.

Third, complete component rectangles are treated as graph four-cycles.

Fourth, the number of four-cycles containing a character is used as a measure of its participation in productive recombination.

Fifth, a directional order parameter measures whether one coordinate carries greater semantic affinity while the other carries greater phonetic affinity.

Sixth, fourth-corner completion tests whether the local geometry predicts withheld linguistic properties.

Seventh, the analysis distinguishes a cycle-rich morphophonetic core from a sparse and irregular lexical periphery.

The principal hypothesis is therefore not simply that component roles exist. It is that repeated recombination converts those roles into a measurable topological geometry.

In compact form:

known linguistic principle + global product representation + graph topology + held-out completion = testable lookup-table law.

The scientific value of the framework will depend on whether it survives changes in decomposition system, pronunciation source, semantic representation, historical period,

character set, and null model. If the same directional structure appears under independent representations, then the writing system contains a robust morphophonetic geometry rather than an artifact of one dataset or annotation scheme.

3. Structural Representation of the Character Inventory

3.1 Character Corpus

The analysis begins with a defined set of Chinese characters for which three kinds of information are available:

1. a standardized character identity,
2. a structural decomposition,
3. linguistic outputs such as pronunciation and meaning.

Let the corpus contain N characters:

C = {c1, c2, ..., cN}.

Each character must have a unique code-point identity, but the numerical code point is treated only as an external label. It is not used as the primary structural representation because code-point values do not directly specify stroke type, component position, or internal composition.

The initial corpus should contain commonly used simplified Chinese characters together with sufficiently many less frequent characters to preserve productive component families. A corpus restricted only to the most frequent characters would be easy to annotate, but it would remove many of the repeated combinations needed to reveal the component lattice. Conversely, a corpus dominated by rare and historical forms might contain too many incomplete or uncertain decompositions.

The working corpus therefore balances several needs:

high-quality structural descriptions,

reliable modern pronunciation,

usable semantic information,

broad coverage of recurring components,

and enough rare combinations to test the boundary between systematic and irregular regions.

Characters with multiple pronunciations or meanings are not discarded. Their outputs are represented as sets or vectors rather than forced into a single label. A character with readings p1 and p2 can be assigned

$P(c) = \{p1, p2\}$.

Similarly, a polysemous character may have several semantic senses:

$M(c) = \{m1, m2, ..., mk\}$.

The corpus also records structural class. At minimum, characters are divided into:

left-right,

top-bottom,

enclosure,

partial enclosure,

single-component,

and complex or uncertain structures.

The distinction is important because a first component and a second component do not always occupy the same physical relation. In a left-right structure, the first operand is usually left of the second. In a top-bottom structure, the first is above the second. In an enclosure structure, the first component may surround the second.

The analysis should preserve this operator information rather than combining all two-component characters into one undifferentiated set.

Variant forms require additional care. Simplified, traditional, historical, and font-specific glyphs may represent related but nonidentical structural objects. The present analysis treats each standardized modern character form as a separate observed entry, while variant and historical relations are stored as external metadata.

The corpus is not assumed to be complete. It is a sampled occupied region of a much larger historical character system. Conclusions must therefore concern the observed character inventory rather than every character ever used.

3.2 Hierarchical Component Decomposition

A Chinese character is not adequately represented by an unordered list of strokes. Its internal arrangement is hierarchical. Components occupy regions, components contain smaller components, and spatial operators determine how they are combined.

A character can be represented as a rooted tree:

T(c).

The leaves are primitive strokes, small graphical elements, or reusable components. Internal nodes are composition operators.

A simple left-right character has the form

T(c) = LR(A, B),

where A is placed on the left and B on the right.

A top-bottom character is

T(c) = TB(A, B).

An enclosing character is

T(c) = ENC(A, B),

where A surrounds or partially surrounds B.

More complicated characters contain nested operations:

T(c) = LR(A, TB(B, C)),

or

T(c) = ENC(A, LR(B, C)).

This representation preserves information that would be lost by flattening the character into a sequence of component names. The expressions

LR(A, B)

and

TB(A, B)

contain the same components but describe different characters because the operator differs.

The decomposition tree also permits analysis at several scales.

At the top level, the character may be divided into two principal components.

At the next level, each principal component may be decomposed into subcomponents.

At the lowest level, the form may be represented as individual strokes or local stroke segments.

This hierarchy permits a direct test of scale. If semantic and phonetic organization is strongest at the level of complete components, then decomposing below that level should weaken the directional separation. If the organization is fundamentally stroke-based, then finer decompositions should preserve or strengthen it.

Let Dk(c) denote the decomposition at depth k.

For example:

D1(c) = top-level components,

D2(c) = subcomponents,

D3(c) = stroke-level elements.

For each scale, the same lookup-table tests can be repeated. This allows the data to determine the privileged structural level rather than assuming in advance that traditional radicals or complete components are optimal.

A decomposition system may assign more than one possible tree to the same character. Rather than concealing this ambiguity, alternative parses can be retained:

T(c) = {T1(c), T2(c), ..., Tr(c)}.

The stability of the lattice result across alternative decompositions becomes an important robustness test.

3.3 Binary Stroke-Position Representation

Component names are useful, but they depend on prior linguistic and lexicographic analysis. To determine whether structure can be recovered from graphical form alone, each character is also represented by an explicit binary stroke-position code.

Let the normalized character square be divided into spatial regions. For every stroke feature type t and region r, define

x(c, t, r) = 1

if character c contains feature type t in region r, and

x(c, t, r) = 0

otherwise.

Flattening all feature-region combinations produces a binary vector

x(c) in {0,1}^D.

Each bit has a fixed interpretation. For example, one bit may indicate:

a horizontal segment in the upper-left region,

a vertical segment in the center,

a hook in the lower-right,

an intersection near the middle,

or an enclosed region occupying the interior.

The binary representation includes more than visible stroke bodies. Separate feature channels may record:

stroke starts,

stroke ends,

bends,

hooks,

crossings,

touching junctions,

enclosure boundaries,

and enclosed interior regions.

A character is therefore represented as a sparse set of active structural features.

The corpus becomes a binary matrix

X in {0,1}^{N x D},

where row c is the structural code for character c.

The matrix is not intended to replace the hierarchical component tree. The two representations answer different questions.

The tree representation identifies reusable symbolic units and composition operators.

The binary representation provides a lower-level geometry that does not require the component names to be known in advance.

If the left and right portions of the binary vector independently recover semantic and phonetic organization, then the directional structure is not merely an artifact of manually assigned component labels.

3.4 Multiscale Spatial Features

A single spatial grid creates a tradeoff.

A coarse grid preserves broad arrangement but cannot distinguish small diagnostic differences.

A fine grid distinguishes local details but becomes sensitive to font variation, stroke displacement, and small normalization errors.

The solution is a multiscale representation.

The character square is encoded at several resolutions, such as

1 x 1,

2 x 2,

4 x 4,

8 x 8,

and 16 x 16.

At the coarsest scale, the code records only which feature types occur anywhere in the character.

At intermediate scales, it records whether they occur in the left, right, upper, lower, central, or corner regions.

At the finest scale, it records local stroke placement.

Let xl(c) be the binary representation at scale l. The full structural code is

x(c) = x1(c) || x2(c) || ... || xL(c),

where || denotes concatenation.

The multiscale code contains intentional redundancy. A stroke represented at fine resolution also contributes to coarser regions. This redundancy improves stability and allows similarity to be measured at several levels simultaneously.

A weighted structural distance between two characters can be defined as

d(c1, c2) = sum over l of alpha_l d_l(c1, c2),

where d_l is the distance at scale l and alpha_l is the scale weight.

For sparse binary vectors, a Jaccard-type distance is preferable to ordinary agreement because shared zeros carry little information:

dJ(c1, c2) = 1 - |A1 intersect A2| / |A1 union A2|,

where A1 and A2 are the active bit sets of the two characters.

The scale weights can initially be fixed without linguistic supervision. Later, they may be optimized separately for component, phonetic, or semantic smoothness. Any such optimization must be evaluated on held-out data to avoid creating an artificial pattern.

3.5 Linguistic Output Tables

The structural matrix X is paired with several output tables. These outputs must remain separate because different parts of the character system may preserve different kinds of information.

The pronunciation table may be represented as

Ysound in {0,1}^{N x P},

where each column corresponds to a pronunciation category. Depending on the analysis, pronunciation may include:

syllable without tone,

syllable with tone,

initial consonant,

rime,

historical reconstruction,

or dialectal reading.

A character with several readings activates several output columns.

The semantic table may be represented as

Ymeaning in R^{N x M}.

Its rows may contain semantic embeddings, dictionary-sense vectors, curated conceptual categories, or membership in semantic classes. Because meaning is more complex than pronunciation, multiple semantic representations should be tested independently.

The component table is

Ycomponent in {0,1}^{N x K},

where each column indicates the presence of a recognized component.

The radical table records dictionary radical or semantic-class assignments.

Additional output tables may include:

stroke count,

frequency,

grammatical category,

historical period,

simplified-traditional relation,

phonetic-series membership,

and lexical usage.

The complete lookup object is therefore not one map but a family of maps:

X -> Ysound,

X -> Ymeaning,

X -> Ycomponent,

X -> Yhistory,

and others.

This separation prevents a strong result in one output domain from being misrepresented as a general result across all domains.

For example, the structural geometry may predict pronunciation strongly, component identity very strongly, and semantics only weakly. These are different findings and should not be merged into one score.

3.6 Separation of Structure and Linguistic Labels

The most important methodological rule is that the structural representation must be constructed without consulting the linguistic outputs that will later be predicted.

If pronunciation or semantic labels influence the structural decomposition, the analysis risks becoming circular.

The structural side should therefore be frozen first.

This includes:

the character corpus,

the composition trees,

the binary stroke-position codes,

the spatial operators,

and the graph edges.

Only after these are fixed should pronunciation, meaning, radical class, and historical labels be attached.

The principle can be written as

construct X without Y,

then test whether X predicts Y.

The same separation applies to parameter selection. If spatial weights are chosen to maximize semantic similarity on all characters, then the resulting semantic structure is partly engineered. A valid procedure divides the data into training and testing sets:

theta_hat = argmax over theta of Score(Xtrain, Ytrain),

followed by

evaluation = Score(Xtest(theta_hat), Ytest).

Null models must preserve relevant structural properties. Randomization should not merely shuffle every bit independently, because that would destroy stroke count, component frequency, and graph degree. Appropriate controls may preserve:

the number of active bits per character,

the frequency of each feature,

stroke-count distribution,

component degrees,

operator class,

and pronunciation frequency.

The purpose is to ask whether the observed alignment between structure and linguistic output exceeds what would arise from simpler properties alone.

By maintaining a strict separation between graphical input and linguistic output, the analysis treats the character inventory as a genuine lookup table. The central question remains open until tested:

Given only the structure of the occupied character space, how much of its semantic and phonological organization can be recovered?

4. The Sparse Morphophonetic Product Lattice

4.1 Two Component Coordinates

For characters with two principal components, the simplest structural model uses two coordinates. Let R denote the first major component and P denote the second. A character is then written as

c = O(R, P),

where O is the spatial composition operator.

For a left-right character, O places R on the left and P on the right. For a top-bottom character, O places R above P. For an enclosure character, O places one component around the other.

The notation does not assume in advance that R is semantic and P is phonetic. At this stage they are only ordered structural coordinates. Their linguistic roles must be established empirically.

Let

R = {r1, r2, ..., rm}

be the set of first-position components, and let

P = {p1, p2, ..., pn}

be the set of second-position components.

The theoretical product table is

R x P.

Each possible pair (ri, pj) defines a potential character position:

cij = O(ri, pj).

Most of these potential combinations are not realized in the observed writing system. Only some pairs correspond to accepted characters.

The occupied character set is therefore

V_O = {(ri, pj) in R x P : O(ri, pj) is an observed character}.

This formulation transforms the inventory from an unordered list into a sparse coordinate system. Every observed character occupies one cell. Characters sharing the same first coordinate lie in the same row. Characters sharing the same second coordinate lie in the same column.

If the first coordinate tends to preserve semantic family and the second tends to preserve pronunciation, then the product table is not merely graphical. It is morphophonetic.

The central question becomes directional:

What linguistic property is preserved when movement occurs along rows?

What linguistic property is preserved when movement occurs along columns?

A row comparison holds the first component fixed while changing the second. A column comparison holds the second component fixed while changing the first.

This gives the two basic relations

same row: O(ri, pj) and O(ri, pk),

same column: O(ri, pj) and O(rh, pj).

If row members are semantically related and column members are phonologically related, then the component coordinates behave as distinct information channels.

4.2 Occupied and Unoccupied Cells

The nominal product table contains

|R| x |P|

possible component pairs.

The observed number of characters is much smaller:

|V_O| << |R| |P|.

Define the occupancy ratio

rho_O = |V_O| / (|R| |P|).

For a highly sparse system, rho_O is close to zero.

This sparsity has several causes. Many component combinations would be graphically awkward, semantically unnecessary, historically unattested, or indistinguishable from existing forms. Some components are productive and combine with many partners, while others occur only once or twice. Historical character formation did not explore the full mathematical product space.

The unoccupied cells should not all be interpreted in the same way.

Some may represent combinations that were never created.

Some may have existed historically but disappeared.

Some may correspond to rare, variant, regional, or obsolete forms missing from the corpus.

Some may be structurally possible but linguistically redundant.

Some may be impossible because the components cannot be placed together under the chosen operator without severe distortion.

The difference between occupied and unoccupied cells is itself informative. The writing system can be studied as a selection process over a much larger possibility space.

A useful distinction is

structurally possible,

historically attested,

currently standardized,

and linguistically productive.

The observed lattice records only one stage of this process.

Sparsity also means that ordinary Cartesian methods may be misleading. The writing system does not fill a smooth rectangular surface. It forms an irregular collection of occupied points connected by shared coordinates.

The correct geometry is therefore partly combinatorial. Distances and neighborhoods should be defined not only by visual similarity but also by paths through observed component recombinations.

4.3 Bipartite Graph Representation

The same product table can be represented as a bipartite graph.

Let the first-coordinate components form one vertex set:

$U = R$.

Let the second-coordinate components form the other:

$W = P$.

Each observed character creates an edge between its two components:

$e_c = (r_i, p_j)$.

The graph is

$G_O = (U, W, E)$,

where E is the set of character edges.

This transformation is useful because many structural properties become standard graph quantities.

The degree of a first-coordinate component r_i is

$\deg(r_i)$ = number of characters containing r_i in the first position.

The degree of a second-coordinate component p_j is

$\deg(p_j)$ = number of characters containing p_j in the second position.

High-degree components are productive. They participate in many characters and create large semantic or phonetic families.

Low-degree components lie near the periphery. They may occur in isolated, archaic, lexicalized, or highly specialized forms.

A path through the graph alternates between components and characters. Two characters may be related because they share one endpoint or because they are connected through a short sequence of recombinations.

The graph representation also prevents a common mistake. If characters are treated only as vertices, component reuse is encoded indirectly. In the bipartite graph, component sharing is explicit: characters are edges, and reusable components are the nodes they connect.

This edge-based view is central to the later four-cycle analysis. A complete 2 x 2 component rectangle corresponds exactly to a cycle of length four in the bipartite graph.

The graph therefore provides a direct bridge between traditional character decomposition and topological analysis.

4.4 Composition Operators

The component pair alone is not sufficient. The same two components can form different characters under different spatial arrangements.

The operator must therefore be included explicitly:

c = O(ri, pj).

At minimum, separate graphs should be constructed for

O = LR,

O = TB,

O = ENC,

and additional operator families when enough data are available.

The left-right graph contains characters of the form

LR(ri, pj).

The top-bottom graph contains

TB(ri, pj).

The enclosure graph contains

ENC(ri, pj).

Combining these without distinction could create false rows and columns. A first component in a left-right structure does not necessarily play the same role as a first component in an enclosure structure.

The operator may also transform component shape. A component compressed into the left side may differ visibly from the same component when placed above or used as an independent character. These contextual variants should be linked to a common abstract component when the relation is clear, while preserving the operator-specific graphical realization.

The general lookup table is therefore not one product space but a union of operator-specific spaces:

V = union over O of V_O.

The associated graph is

G = disjoint or linked union over O of G_O.

The role of each coordinate may depend on O. In the dominant left-right system, the first coordinate may often be semantic and the second phonetic. In top-bottom or enclosure structures, the same functional roles may occupy different spatial directions.

This produces the operator-covariance question:

Does the semantic-phonetic distinction remain attached to operand order, or is it attached to physical position?

For example, if the semantic component is first under both LR and TB, then the functional axis follows operand order. If it is usually left under LR but above or outside under other operators, then the geometry rotates with the operator.

This question cannot be answered by analyzing left-right characters alone.

4.5 Connectivity and the Giant Component

A sparse graph may still be highly connected.

Let G_O contain several connected components. A connected component is a maximal set of component nodes linked through observed character edges.

If most productive components belong to one very large connected component, then the character system has a giant recombinational core. Components are not confined to isolated families. They participate in chains of shared combinations that connect large portions of the inventory.

Let

N_total = number of component vertices,

N_giant = number of vertices in the largest connected component.

Define the giant-component fraction

g = N_giant / N_total.

A value of g near 1 indicates that most components belong to one connected structural network.

Connectivity does not imply linguistic uniformity. Two distant components in the same graph may have no direct semantic or phonetic relation. However, connectivity means that they participate in one extended system of recombination.

The giant component provides the domain in which large-scale topology can be studied. Within it, one can identify:

dense regions,

bridges,

articulation points,

high-degree hubs,

cycle-rich neighborhoods,

and sparse branches.

Isolated components and small connected components form a different regime. They may represent rare structures, incomplete decomposition, exceptional historical forms, or genuinely independent character families.

The distinction between giant core and isolated periphery is therefore not imposed from outside. It emerges from the component-reuse graph itself.

4.6 Productive Core and Sparse Periphery

Degree alone does not fully define productivity. A component may have high degree because it combines with many partners, but the resulting subgraph may still lack repeated rectangular structure.

A stronger measure asks whether components participate in recombinational closure.

Suppose ri combines with pj and pk, while rh also combines with pj. If rh additionally combines with pk, the four combinations form a complete rectangle.

Such closure indicates that two first-coordinate components and two second-coordinate components have been systematically recombined.

The productive core is therefore characterized by more than frequency. It contains many overlapping component rectangles and many short cycles.

The sparse periphery contains characters with few or no such cycles.

A character in the core can be compared with multiple analogues:

same first component, different second component,

same second component, different first component,

and complete three-corner supports for fourth-corner prediction.

A peripheral character may have no comparable neighbors.

This motivates a topological classification.

Let q(c) be the number of complete four-cycles containing character c.

Then

q(c) = 0

means the character participates in no complete component rectangle.

A large value of q(c) means the character belongs to many overlapping local product cells.

The hypothesis is that linguistic regularity increases with q.

Characters in the high-q core should more often obey the expected semantic and phonetic roles.

Characters in the low-q periphery should contain more exceptions, role reversals, historical opacity, lexicalized meanings, and pronunciation divergence.

The character inventory may therefore contain two broad regimes:

an orderly morphophonetic core,

and an irregular lexical periphery.

The boundary between them need not be sharp in every dataset. It can be studied by examining how semantic and phonetic order measures change as functions of degree, cycle count, and local density.

This converts the vague idea of productive character formation into a measurable graph property.

The principal claim of the product-lattice model is therefore not that every character lies in a perfect table. It is that repeated component recombination creates local regions that behave like a product code, and that the strength of this behavior can be measured through the topology of the occupied lattice.

5. Dual Semantic and Phonetic Channels

5.1 First-Component Semantic Affinity

The product-lattice hypothesis predicts that the first major component should often preserve semantic information. This prediction is strongest in structural classes where the first operand commonly functions as a semantic classifier, especially the large left-right character family.

Let a character be written as

$c = O(R, P)$,

where R is the first component, P is the second component, and O is the composition operator.

For two characters sharing the first component,

$c1 = O(R, P1)$

and

$c2 = O(R, P2)$,

the semantic-channel hypothesis predicts

$Sim_meaning(c1, c2) > Sim_meaning(\text{random matched pair})$.

The comparison must be made against controls matched for factors such as stroke count, frequency, character complexity, and component productivity. Otherwise, frequently occurring components might appear semantically informative merely because they occur in many common characters.

Semantic similarity may be measured in several ways. One method compares dictionary glosses. Another places characters in a vector space derived from Chinese-language usage. A third uses curated semantic classes or lexical resources. Because each method has limitations, the strongest conclusion comes from agreement across several independent semantic representations.

The semantic component is not expected to specify the exact lexical meaning. It usually identifies a broad domain. A component associated with water may occur in characters referring to rivers, washing, fluids, moisture, depth, or movement through water. The shared information is therefore categorical or field-like rather than definitional.

The appropriate question is not

Does R determine M(c)?

but

Does knowing R reduce uncertainty about M(c)?

This can be expressed using conditional entropy:

$H(M \mid R) < H(M)$.

The information supplied by the first component is

$I(R; M) = H(M) - H(M \mid R)$.

A positive value means that first-component identity reduces uncertainty about semantic class.

A local similarity measure can also be defined. Let $N_R(c)$ be the set of characters sharing the first component with c. Then

$S_R(c)$ = average over d in $N_R(c)$ of Sim_meaning(c, d).

The corpus-level semantic affinity is

S_R = average over c of $S_R(c)$.

This quantity is compared with null neighborhoods having similar size and structural complexity.

The expected result is not perfect semantic agreement. The semantic component may have become opaque, generalized, narrowed, or historically detached from the modern meaning. Nevertheless, a statistically elevated semantic affinity indicates that the first coordinate carries semantic information across the lookup table.

5.2 Second-Component Phonetic Affinity

The second component is tested in the same way, but the target output is pronunciation.

For two characters sharing the second component,

c1 = O(R1, P)

and

c2 = O(R2, P),

the phonetic-channel hypothesis predicts

Sim_sound(c1, c2) > Sim_sound(random matched pair).

Pronunciation similarity can be defined at several levels.

The strictest measure asks whether the complete modern reading is identical, including tone.

A broader measure ignores tone and compares the syllable base.

More detailed measures separately compare:

initial consonant,

medial,

vowel nucleus,

final consonant,

tone,

and reconstructed historical reading.

A binary same-pinyin measure is useful but incomplete. Two readings may differ by one sound and still belong to the same historical phonetic series. A graded phonological distance is therefore preferable for some analyses.

Let p(c) be a phonological feature vector. A simple distance is

d_sound(c1, c2) = sum over k of w_k |p_k(c1) - p_k(c2)|,

where the weights w_k specify the relative importance of initials, finals, tones, and other features.

The corresponding similarity may be

Sim_sound(c1, c2) = exp(-d_sound(c1, c2)).

For each character c, let N_P(c) be the set of characters sharing its second component. The phonetic affinity is

S_P(c) = average over d in N_P(c) of Sim_sound(c, d).

The full-corpus value is

S_P = average over c of S_P(c).

The second-component channel may be stronger than the first-component semantic channel because pronunciation families are often more sharply defined than semantic fields. Two characters sharing a phonetic component may have unrelated meanings while retaining related readings.

Modern pronunciation is not the only relevant target. Chinese sound systems have changed substantially over time. A component that appears irregular in present-day Mandarin may have been highly predictive when the character was created. Therefore, historical reconstructions and major dialect readings provide important secondary tests.

The phonetic component should be understood as a probabilistic transmitter through historical change. Its original output has been altered by sound shifts, tone development, dialect divergence, borrowing, and character-specific exceptions.

Even with these complications, the prediction remains direct:

P(similar pronunciation | shared second component)

should greatly exceed

P(similar pronunciation | matched random pair).

5.3 Role Reversal Controls

A two-channel model requires more than showing that both components contain information. It requires directional specialization.

The first component should be more semantically informative than the second, while the second should be more phonologically informative than the first.

Define

Sem1 = semantic affinity through the first component,

Sem2 = semantic affinity through the second component,

Pho1 = phonetic affinity through the first component,

Pho2 = phonetic affinity through the second component.

The expected inequalities are

Sem1 > Sem2

and

Pho2 > Pho1.

A combined directional score can be defined as

D = (Sem1 - Sem2) + (Pho2 - Pho1).

A positive D indicates the expected semantic-first, phonetic-second orientation.

This role-reversal control is essential because visual components can correlate with many outputs for indirect reasons. For example, a productive phonetic component may occur mostly within one broad semantic class, creating apparent semantic similarity. Likewise, a semantic component may occur in a historical pronunciation cluster.

The directional test asks whether each coordinate carries its predicted output more strongly than the alternative output.

A stronger control swaps the labels assigned to the two coordinates. The model is evaluated under two hypotheses:

H1: first coordinate is semantic and second is phonetic,

H2: first coordinate is phonetic and second is semantic.

The observed system supports H1 only if its predictive performance exceeds H2 under the same data, scoring rules, and null models.

The comparison should also be made within characters rather than only across separate groups. For every eligible character, one can estimate how well its first-component neighbors predict its meaning and how well its second-component neighbors predict its pronunciation.

This paired design reduces the influence of character frequency, complexity, and annotation quality.

The resulting directional asymmetry is the central evidence for dual channels. Without it, the table might contain only general component clustering rather than a morphophonetic product geometry.

5.4 Spatial Recovery Without Component Names

Component decompositions contain prior knowledge. If a database already labels one part as the first component and another as the second, a critic may argue that the dual-channel result was built into the representation.

A stronger test begins with raw graphical structure.

Each character is represented by a binary stroke-position vector

$x(c) \in \{0,1\}^D$.

The spatial code is partitioned into complementary regions. For left-right characters, define

$x_L(c)$ = features in the left region,

$x_R(c)$ = features in the right region.

The question is whether these unnamed spatial channels reproduce the linguistic asymmetry:

left-side similarity predicts meaning,

right-side similarity predicts pronunciation.

For two characters c1 and c2, define

J_L(c1, c2) = weighted Jaccard similarity of x_L(c1) and x_L(c2),

J_R(c1, c2) = weighted Jaccard similarity of x_R(c1) and x_R(c2).

The left semantic test asks whether high J_L predicts semantic similarity better than high J_R.

The right phonetic test asks whether high J_R predicts pronunciation similarity better than high J_L.

If the expected asymmetry appears without component labels, then the result is present in graphical geometry itself.

Alternative partitions should be tested:

left versus right,

top versus bottom,

outer versus inner,

central versus peripheral,

and diagonal or learned partitions.

The preferred partition is the one maximizing directional separation on training data and retaining that advantage on held-out characters.

A general spatial score is

D_spatial(theta) =
Sem_A(theta) - Sem_B(theta)

- Pho_B(theta) - Pho_A(theta),

where theta defines a partition of the character square into regions A and B.

For left-right characters, the hypothesis predicts a maximum near a vertical division.

For top-bottom characters, the maximum may rotate toward a horizontal division.

For enclosure structures, the natural split may be outer versus inner rather than left versus right.

The spatial recovery test serves two purposes. It validates the component analysis independently, and it reveals whether the linguistic channels correspond to physical location, operand order, or both.

5.5 The Privileged Decomposition Scale

A character can be decomposed at several levels. The semantic-phonetic separation may operate at one level more strongly than at others.

Consider three scales:

top-level components,

smaller subcomponents,

individual strokes or local stroke features.

At the top level,

$c = O(A, B)$.

At a deeper level,

$A = O1(a1, a2)$

and

$B = O2(b1, b2)$.

At the stroke level, the character is represented as a large collection of primitive features.

The question is whether the directional score D is maximal at one particular scale.

Let D_k be the semantic-phonetic separation obtained from decomposition level k.

A privileged component scale exists if

$D_{top} > D_{subcomponent}$

and

$D_{top} > D_{stroke}$.

This result would show that the relevant units are not arbitrary collections of strokes. Complete components would function as stable information-bearing tokens.

At excessively coarse resolution, the character is treated as one indivisible form, and the two channels cannot be separated.

At excessively fine resolution, meaningful components are fragmented into strokes that individually carry little semantic or phonological information.

The expected pattern is therefore nonmonotonic:

weak separation at whole-character level,

strong separation at major-component level,

weaker separation at subcomponent and stroke levels.

The preferred scale may be identified by maximizing

D_k - lambda C_k,

where C_k is representational complexity and lambda penalizes unnecessary detail.

A minimum-description-length formulation can also be used. The best scale is the one that explains the greatest amount of semantic and phonetic structure using the fewest reusable units.

The existence of a privileged scale would strengthen the analogy with a code. It would identify the natural symbol size at which the lookup table becomes orderly.

5.6 Operator-Dependent Orientation

The dual-channel model must be tested separately for each composition operator.

For left-right characters,

c = LR(A, B),

the common hypothesis is

A -> semantic tendency,

B -> phonetic tendency.

For top-bottom characters,

c = TB(A, B),

the corresponding functions may be attached to upper and lower position, to operand order, or to specific component families.

For enclosure characters,

c = ENC(A, B),

the relevant distinction is often outer versus inner.

Three competing models can be defined.

The position-fixed model states that semantic and phonetic functions remain attached to physical directions such as left and right.

The operand-fixed model states that the first operand tends to be semantic and the second tends to be phonetic regardless of spatial orientation.

The operator-specific model states that every operator has its own role distribution:

roles = F(O, A, B).

The data can distinguish among these possibilities.

For each operator O, define

D_O = (Sem1_O - Sem2_O) + (Pho2_O - Pho1_O).

If D_O is positive across several operators, then the dual-channel distinction follows operand order.

If only left-right characters show a strong positive value, then the pattern may be specific to that layout.

If different operators have different signs or preferred partitions, then the morphophonetic axes rotate with structural composition.

Operator-specific analysis is necessary because pooling all layouts can cancel real effects. A semantic component above a phonetic component and a semantic component outside a phonetic component may be functionally equivalent but spatially different.

The general model is therefore

c = O(R, P) + epsilon_O,

where epsilon_O includes operator-specific distortion and exceptions.

The strongest version of the hypothesis predicts covariance:

when the composition operator changes, the physical orientation of the two channels changes, but their linguistic functions remain distinguishable.

Establishing such covariance would show that the product geometry is not merely a left-right visual pattern. It would be a structural rule governing how reusable components transmit semantic and phonological information across multiple character architectures.

6. Four-Cycle Topology and the Morphophonetic Order Parameter

6.1 Complete Character Rectangles

The sparse product table becomes mathematically informative when component pairs are reused in complete 2 x 2 patterns. Suppose two first-coordinate components, R1 and R2, combine with two second-coordinate components, P1 and P2. If all four combinations exist, then the table contains

C11 = O(R1, P1)

C12 = O(R1, P2)

C21 = O(R2, P1)

C22 = O(R2, P2).

These four characters form a complete rectangle in the component table.

In the corresponding bipartite graph, R1 and R2 are vertices on one side, P1 and P2 are vertices on the other, and the four characters are edges. The four edges create a closed path of length four:

R1 -> P1 -> R2 -> P2 -> R1.

This is a graph four-cycle.

The rectangle is important because it separates the two component coordinates internally. The two characters in a row share the same first component:

C11 and C12 share R1,

C21 and C22 share R2.

The two characters in a column share the same second component:

C11 and C21 share P1,

C12 and C22 share P2.

If the first coordinate tends to preserve semantic family, then row pairs should be semantically more similar than column pairs. If the second coordinate tends to preserve pronunciation, then column pairs should be phonologically more similar than row pairs.

A complete rectangle therefore acts as a self-controlled local experiment. The same four characters are used to compare both directions. Differences in character frequency, complexity,

historical period, and annotation quality are partly controlled because the row and column tests operate within the same motif.

The rectangle is the smallest product structure capable of distinguishing two functional axes. A single edge shows only one character. Two edges sharing a component show one family. Three corners suggest a possible analogy. Four occupied corners establish a closed local product cell.

The central hypothesis is that these cells are not linguistically neutral. They should exhibit directional organization:

row direction -> semantic continuity,

column direction -> phonetic continuity.

6.2 Four-Cycle Degree

Not all characters participate equally in complete rectangles.

For a character c represented by edge e_c, define q(c) as the number of distinct four-cycles containing that edge.

A character with

q(c) = 0

belongs to no complete rectangle. It may share one component with other characters, but it lacks a closed two-coordinate comparison.

A character with

q(c) = 1

belongs to one local product cell.

A character with large q belongs to many overlapping rectangles and occupies a highly recombinatorial region of the component graph.

The value q measures something different from ordinary component frequency.

A character may contain two high-frequency components but still participate in few complete rectangles if those components do not recombine systematically with the same partners.

Conversely, a moderately frequent character may participate in many cycles if its components lie within a densely interwoven product family.

The degree of an edge’s endpoints can be written as

d_R(c) = degree of its first component,

d_P(c) = degree of its second component.

These quantities measure component productivity. The four-cycle degree q(c) measures local closure.

A rough upper bound on cycle participation is

q(c) <= (d_R(c) - 1)(d_P(c) - 1).

Equality would occur only if every alternative partner of the first component also combines with every alternative partner of the second component.

A normalized closure score can therefore be defined as

kappa(c) = q(c) / [(d_R(c) - 1)(d_P(c) - 1)],

when both endpoint degrees exceed 1.

The raw count q measures how many analogical squares surround the character. The normalized value kappa measures how completely the local product possibilities are realized.

Both quantities may be informative.

High q indicates a large supply of analogical evidence.

High kappa indicates strong local combinatorial closure.

A character can have high q but modest kappa if it lies in a very large but only partly filled network. Another may have low q but high kappa if it lies in a small, nearly complete family.

The main empirical question is whether morphophonetic regularity increases with q, kappa, or both.

6.3 Morphophonetic Order Parameter

To quantify directional organization, define four local affinities for each eligible character or rectangle.

Sem1 measures semantic similarity among characters sharing the first component.

Sem2 measures semantic similarity among characters sharing the second component.

Pho1 measures phonetic similarity among characters sharing the first component.

Pho2 measures phonetic similarity among characters sharing the second component.

The expected morphophonetic orientation is

Sem1 > Sem2

and

Pho2 > Pho1.

A combined order parameter is

M = (Sem1 - Sem2) + (Pho2 - Pho1).

A positive M indicates that the first coordinate is more semantic and the second more phonetic.

A negative M indicates reversal or disorder.

A value near zero indicates weak differentiation between the coordinates.

The two terms can also be weighted:

M = alpha(Sem1 - Sem2) + beta(Pho2 - Pho1),

where alpha and beta compensate for differences in measurement scale or reliability.

To avoid arbitrary weighting, semantic and phonetic scores can first be standardized relative to their null distributions:

Z_sem = [Sem1 - Sem2 - mu_sem,null] / sigma_sem,null

Z_pho = [Pho2 - Pho1 - mu_pho,null] / sigma_pho,null.

Then

M_Z = Z_sem + Z_pho.

The order parameter may be calculated at several levels:

for individual characters,

for individual rectangles,

for component families,

for operator classes,

and for bins of cycle degree.

At the rectangle level, one possible semantic term is

Sem_rect =
0.5[SimM(C11,C12) + SimM(C21,C22)]

- 0.5[SimM(C11,C21) + SimM(C12,C22)].

The phonetic term is

Pho_rect =
0.5[SimP(C11,C21) + SimP(C12,C22)]

- 0.5[SimP(C11,C12) + SimP(C21,C22)].

The rectangle order is

M_rect = Sem_rect + Pho_rect.

This score directly compares the two orientations inside one four-character square.

A highly ordered rectangle has positive semantic continuity across rows and positive phonetic continuity down columns.

6.4 Transition With Cycle Density

The strongest topological hypothesis is not merely that four-cycles exist. It is that increasing participation in four-cycles produces increasing morphophonetic order.

Characters can be divided into cycle bins such as

q = 0,

q = 1 to 5,

q = 6 to 20,

q = 21 to 100,

q > 100.

For each bin, calculate the mean order parameter:

M_bar(q-bin).

If the productive-core hypothesis is correct, then

M_bar(high q) > M_bar(low q).

The relation need not be linear. A threshold-like transition may occur. Sparse regions may show little or no directional order until enough alternative component combinations exist to stabilize the semantic and phonetic roles.

A possible phenomenological model is

M(q) = M0 + A q^gamma / (q^gamma + q0^gamma),

where M0 is the peripheral baseline, A is the increase in the ordered core, q0 is the characteristic transition scale, and gamma controls the sharpness of the transition.

This form is not assumed to be fundamental. It provides a way to estimate whether the system displays gradual saturation or a sharp change.

A simpler model is

$M(q) = a + b \log(1 + q)$.

A positive b indicates increasing order with cycle participation.

The transition should be examined separately for each operator class. The characteristic q0 may differ between left-right, top-bottom, and enclosure graphs because their sizes and component productivities differ.

A strong result would show that:

low-cycle characters are weakly ordered or disordered,

intermediate-cycle characters show mixed behavior,

high-cycle characters show stable semantic-first and phonetic-second organization.

Such a transition would suggest that repeated recombination does not merely reflect morphophonetic order. It may help define the region in which the order becomes operationally reliable.

6.5 Degree-Controlled Null Models

Cycle-rich characters tend to contain high-degree components. Therefore, an observed relation between q and M could arise simply because frequent components are better annotated, more regular, or more likely to belong to familiar families.

The analysis must separate cycle closure from endpoint degree.

A regression model can be written as

$M(c) = a$

- $b_q \log(1 + q(c))$
- $b_R \log(1 + d_R(c))$
- $b_P \log(1 + d_P(c))$
- $b_f \log(1 + \text{frequency}(c))$
- $\text{epsilon}(c)$.

The key coefficient is b_q.

If b_q remains positive after controlling for both component degrees and character frequency, then four-cycle participation contributes information beyond component productivity alone.

A nonparametric alternative constructs matched controls. For each character c, select characters with similar

d_R,

d_P,

stroke count,

frequency,

operator class,

and component complexity,

but different q.

The difference in M between high-q and low-q matched characters estimates the effect of cycle closure more directly.

A graph-rewiring null is also useful. The bipartite graph can be randomized while preserving every component degree. Degree-preserving edge swaps maintain the number of characters associated with each component but alter which pairs occur and therefore alter the four-cycle structure.

For every randomized graph, recalculate the cycle counts and linguistic order scores. The observed result is significant if the real graph produces stronger alignment between q and M than the rewired ensemble.

Another null preserves the graph exactly but permutes semantic and phonetic labels among character edges. This tests whether the observed topology is aligned with linguistic outputs rather than merely structurally unusual.

A particularly strict control permutes labels only within matched groups defined by stroke count, frequency, and component degree. This preserves broad distributional effects while destroying the specific coordinate-output relationship.

The purpose of these controls is to establish that the order parameter is linked to topological recombination, not merely to frequent characters or large families.

6.6 A Topological Phase Interpretation

If M rises sharply with q and remains associated with q after degree controls, then the character graph may be described as containing two regimes.

The first is a sparse lexical periphery.

Characters in this regime participate in few or no complete rectangles. Their components may be unique, historically opaque, irregularly assigned, or weakly productive. Semantic and phonetic roles may be reversed, mixed, or uninformative.

The second is a cycle-rich morphophonetic core.

Characters in this regime participate in many overlapping rectangles. Their components are repeatedly recombined across semantic and phonetic families. The first and second coordinates behave more consistently as separate information channels.

The term phase is used here in a graph-theoretic and phenomenological sense. It does not imply a literal thermodynamic transition. It refers to a qualitative change in collective order as local recombination increases.

The proposed order parameter is M.

The proposed structural control variable is q or kappa.

The ordered regime is characterized by

$M \gg 0$.

The disordered or mixed regime is characterized by

M approximately 0

or

$M < 0$.

The analogy to a phase transition becomes more convincing if several properties change together near the same cycle-density range:

semantic row affinity increases,

phonetic column affinity increases,

fourth-corner prediction improves,

exceptions decline,

and operator-specific orientation stabilizes.

The topological interpretation also explains why global analyses can appear weak. The full corpus combines an orderly core with a large irregular periphery. Averaging across both regimes dilutes the signal.

The relevant law may therefore be conditional:

Morphophonetic product order is strong where the component graph contains sufficient four-cycle closure.

This is a more precise claim than saying that Chinese characters are generally semantic-phonetic compounds. It predicts where the principle should work best, where it should fail, and which graph property distinguishes the two regions.

The four-cycle framework thus converts traditional component analysis into a testable theory of collective structure. Complete rectangles become the elementary cells, cycle count becomes a measure of local participation in the product code, and the morphophonetic order parameter measures the directional specialization of the two component coordinates.

7. Fourth-Corner Completion

7.1 Definition of the Fourth Corner

A complete component rectangle contains four characters:

C11 = O(R1, P1)

C12 = O(R1, P2)

C21 = O(R2, P1)

C22 = O(R2, P2).

The first coordinate is shared across rows, and the second coordinate is shared across columns. If one member is hidden, the remaining three define an analogical prediction problem.

Suppose C22 is withheld. The visible characters are

C11 = O(R1, P1)

C12 = O(R1, P2)

C21 = O(R2, P1).

The missing character must occupy the cell

C22 = O(R2, P2).

This unknown cell is called the fourth corner. In classical analogy, the same relation is written

A : B :: C : X,

where X is the fourth proportional.

The term fourth corner emphasizes the geometry of the 2 x 2 table. Three occupied corners define two coordinate directions. The missing corner is predicted by combining the second row coordinate with the second column coordinate.

In English morphology, the pattern

walk : walked :: talk : X

gives

X = talked.

In the Chinese component lattice, the prediction is not only graphical. The row and column are expected to transmit different kinds of linguistic information.

The row coordinate predicts semantic family.

The column coordinate predicts pronunciation family.

Fourth-corner completion therefore asks whether the local geometry functions as a decoding rule.

7.2 Chinese Semantic Prediction

For the hidden character C22, the known row peer is C21. Both share the first component R2:

C21 = O(R2, P1)

C22 = O(R2, P2).

If R2 carries semantic information, then the meaning of C21 should help predict the semantic field of C22.

The simplest semantic predictor is

M_hat(C22) = M(C21),

where M is a semantic representation.

This does not imply that the two characters have identical meanings. The prediction concerns semantic family or similarity.

A more general prediction combines all visible or training characters sharing R2:

M_hat(C22) =
average over c with first component R2 of M(c),

excluding C22 itself.

If semantic vectors are normalized, prediction quality can be measured by cosine similarity:

Score_sem(C22) =
cosine[M_hat(C22), M(C22)].

For categorical semantic labels, the predictor may use majority class, weighted voting, or probability estimates.

The row prediction should be compared with several baselines:

the global mean semantic vector,

a random character of similar frequency,

a character matched by stroke count,

a character sharing only the second component,

and a character selected from a degree-matched randomized graph.

The central semantic comparison is between the correct row axis and the swapped column axis:

Score_sem,row > Score_sem,column.

Because semantic components often indicate broad categories, semantic prediction is expected to be graded rather than exact. A water-related row may contain characters involving rivers, washing, moisture, waves, or liquids. The row signal may therefore be strong at the category level while weak at the level of exact definition.

7.3 Chinese Pronunciation Prediction

For the hidden character C22, the known column peer is C12. Both share the second component P2:

C12 = O(R1, P2)

C22 = O(R2, P2).

If P2 carries phonetic information, then the pronunciation of C12 should help predict the pronunciation of C22.

The simplest predictor is

P_hat(C22) = P(C12),

where P is the pronunciation label.

When characters have multiple readings, prediction may be set-valued:

P_hat(C22) = set of readings observed among characters sharing P2.

A majority-vote predictor uses all training characters with second component P2:

P_hat(C22) =
mode of P(c) over c sharing P2,

excluding C22.

Pronunciation can be scored at several levels:

exact syllable and tone,

syllable without tone,

initial consonant,

final,

historical reconstruction,

or graded phonological distance.

A binary exact-match score is

Hit(C22) =
1 if P_hat(C22) contains a true reading of C22,
0 otherwise.

A graded score may be defined as

Score_pho(C22) =
1 - d_pho[P_hat(C22), P(C22)],

where d_pho is a normalized phonological distance.

The decisive comparison is

Score_pho,column > Score_pho,row.

If the column prediction is strong while the row prediction is near chance, the two coordinates are functionally specialized.

Pronunciation completion is often expected to outperform semantic completion because phonetic series can preserve relatively specific syllabic patterns, whereas semantic components usually provide only broad conceptual guidance.

7.4 Strict Three-Corner Test

The strictest version of fourth-corner completion uses only one rectangle.

For target C22, the algorithm receives

C11,

C12,

C21,

but not the linguistic labels of C22.

The supporting rectangle must be selected without examining the hidden target output. Otherwise, selection could favor a square known in advance to succeed.

A valid selection rule may use only structural information, such as:

component degrees,

cycle count,

character frequency,

graph distance,

or graphical similarity among the three visible corners.

Once the rectangle is chosen, semantic prediction comes from C21 and phonetic prediction comes from C12.

The strict test is valuable because it closely matches the intuitive idea of analogy. Only three known examples are used.

Its weakness is sensitivity to noise. One row peer may be semantically atypical, and one column peer may have an irregular modern pronunciation.

The strict test therefore measures the strength of individual local analogies, not the maximum information available in the full graph.

For each target, the reported outputs should include

the selected rectangle,

the three visible corners,

the predicted semantic field,

the predicted pronunciation,

the actual hidden outputs,

and whether the completion succeeded.

Transparent examples are especially important because they show that the score is produced by an understandable local rule rather than an opaque statistical model.

7.5 Aggregated Local Completion

A character in the cycle-rich core may belong to many rectangles. Instead of choosing one, the aggregated predictor combines evidence from all supporting squares.

Let S_row(C22) be the set of row peers connected to C22 through complete rectangles. Let S_col(C22) be the corresponding set of column peers.

The semantic prediction is

M_hat(C22) =
sum over c in S_row(C22) of w_c M(c)
divided by
sum over c in S_row(C22) of w_c.

The pronunciation prediction is obtained from weighted voting over S_col(C22).

Weights may depend on

the cycle density of the supporting rectangle,

component frequency,

historical reliability,

graph similarity,

or agreement among the three known corners.

A simple choice is uniform weighting:

w_c = 1.

A more selective choice gives greater weight to supports from highly ordered rectangles:

w_c = max(0, M_rect).

Aggregated completion should outperform strict completion when multiple independent analogies agree. It should also provide a confidence estimate.

For pronunciation, confidence may be

Conf_pho =
votes for leading pronunciation
divided by
total votes.

For semantics, confidence may be based on the concentration of the predicted semantic vectors.

The aggregated method tests whether the graph behaves as a redundant error-correcting system. Multiple rectangles provide repeated evidence for the same hidden output.

However, the strict and aggregated results must both be reported. Strong aggregated performance with weak strict performance would show that prediction depends on pooling many weak signals. Strong performance in both would indicate that individual squares themselves function as useful decoding circuits.

7.6 Swapped-Axis and Randomized Controls

Fourth-corner completion is meaningful only if the correct coordinate directions outperform plausible alternatives.

The first control swaps the axes.

Semantic prediction is made from the column peer instead of the row peer.

Pronunciation prediction is made from the row peer instead of the column peer.

The expected inequalities are

Semantic_correct > Semantic_swapped

and

Phonetic_correct > Phonetic_swapped.

The second control permutes linguistic labels while preserving the graph. This retains the same component structure and cycle counts but destroys the alignment between topology and linguistic output.

The third control rewires the bipartite graph while preserving component degrees. This keeps component productivity constant but changes which component pairs form characters and which rectangles exist.

The fourth control matches targets by stroke count, frequency, operator class, and component degree.

The fifth control replaces complete rectangles with open three-edge paths that do not close into a fourth observed character. If closed cycles carry special information, completion from genuine rectangles should outperform completion from structurally similar but unclosed motifs.

A useful effect size for semantics is

Delta_sem =
Score_sem,correct - Score_sem,swapped.

For pronunciation,

Delta_pho =
Score_pho,correct - Score_pho,swapped.

The combined completion order is

Delta_total = Delta_sem + Delta_pho.

Permutation tests estimate how often a randomized system produces a value at least as large as the observed one.

These controls distinguish a true directional lookup law from general component similarity, frequency effects, or accidental clustering.

7.7 Representative Character Squares

A transparent example can be constructed from two semantic-side components and two phonetic-side components.

Consider the following rectangle:

	P1	P2
R1	C11	C12
R2	C21	C22

Suppose R1 is associated with water-related meanings and R2 with fish-related meanings. Suppose P1 suggests the pronunciation xun and P2 suggests huang.

The observed characters may then behave schematically as

water + P1 -> water-related character pronounced xun

water + P2 -> water-related character pronounced huang

fish + P1 -> fish-related character pronounced xun

fish + P2 -> fish-related character pronounced huang.

If the final character is hidden, the other three corners predict

semantic family: fish,

pronunciation family: huang.

The exact lexical meaning is not determined, but the two major output channels are constrained.

Such examples illustrate the product interpretation:

moving horizontally changes P while holding R fixed,

moving vertically changes R while holding P fixed.

A good set of examples should include

high-confidence successful squares,

partially successful squares,

historically irregular pronunciations,

semantic drift,

and failures.

Failures are scientifically important. They reveal where the local product law breaks down and may identify historical changes or incorrect decompositions.

7.8 Completion Accuracy as Evidence of a Lookup Law

Clustering shows that related characters tend to occur near one another. Correlation shows that structural similarity is associated with linguistic similarity. Fourth-corner completion is stronger because it requires prediction of withheld information.

The model is not asked whether known row pairs are semantically similar after all labels have been observed. It is asked to estimate the hidden character's semantic field without using that field.

Likewise, it must predict pronunciation from the surrounding column structure without seeing the target reading.

A successful completion law should satisfy several conditions.

First, prediction must be out of sample. The target's labels cannot contribute to its own predictor.

Second, the supporting rectangle must be selected without consulting the hidden outputs.

Third, the correct axes must outperform swapped axes.

Fourth, performance must exceed matched and randomized controls.

Fifth, accuracy should increase with four-cycle degree if cycle-rich regions truly form the ordered core.

Sixth, the result should replicate across independent decomposition and linguistic datasets.

The central prediction is

Accuracy(q high) > Accuracy(q low).

This relation connects topology directly to function.

If high-cycle characters show strong fourth-corner completion while low-cycle characters do not, then the cycle count acts as an order parameter for predictive regularity.

The resulting law is local rather than universal:

Within the cycle-rich morphophonetic core, three occupied corners of a component rectangle constrain the linguistic properties of the fourth.

This is the most direct expression of the proposed hidden geometry. The writing system is not merely composed of reusable parts. In its densely recombined regions, those parts form local product cells that transmit semantic and phonetic information along different directions.

8. Nonlinear Geometry of the Character System

8.1 Limits of a Global Linear Map

A simple mathematical approach treats the character structure as a binary vector x and attempts to predict linguistic outputs through a linear map:

Y_hat = XA.

Here X is the character-by-feature matrix, A is a coefficient matrix, and Y contains semantic, phonological, or component labels.

This model assumes that each structural feature contributes independently and additively. A horizontal stroke in one region receives one coefficient, an enclosure feature receives another, and the output is the sum of all such contributions.

That assumption is too restrictive for Chinese character structure.

The effect of a graphical feature usually depends on its context. A stroke pattern may function differently depending on:

its position,

the component containing it,

the component beside it,

the composition operator,

and the larger character family.

A feature that is informative on the left side of a left-right character may be uninformative or carry a different role when placed at the bottom of a top-bottom character.

Likewise, the same component may function semantically in one family and phonetically in another.

The true mapping therefore cannot generally be written as

Y = sum over i of a_i x_i.

A more realistic expression includes interactions:

Y = sum over i of a_i x_i

- sum over i,j of b_ij x_i x_j
- sum over i,j,k of c_ijk x_i x_j x_k
- higher-order terms.

The interaction term x_i x_j becomes active only when both features are present. This captures the idea that a combination may carry information not present in either feature alone.

However, even a large polynomial expansion may be conceptually inadequate. The important nonlinearities are not arbitrary interactions among thousands of bits. They are structured interactions among components, positions, and composition operators.

The weak performance of a global linear model therefore does not imply that the lookup table lacks order. It may imply that the wrong mathematical form is being used.

A single linear map attempts to fit one plane through a system that may contain several curved, branching, or piecewise-regular regions.

8.2 Conditional Component Functions

The component model suggests that the linguistic output should be written as a conditional function:

Y = F(R, P, O, epsilon),

where R is the first major component, P is the second, O is the composition operator, and epsilon represents historical and graphical residuals.

The contribution of R is not independent of P.

The contribution of P is not independent of R.

The role of both depends on O.

For example, the phonetic value of a component may be reliable only when it appears in a particular position or with a particular class of semantic components.

The semantic contribution of a radical-like component may also depend on the lexical domain of the accompanying phonetic component.

The mapping is therefore relational rather than additive.

For each operator, one may define a separate function:

Y_LR = F_LR(R, P)

Y_TB = F_TB(R, P)

Y_ENC = F_ENC(R, P).

These functions need not have the same parameters or geometry.

Within the left-right system, the semantic and phonetic channels may be relatively stable:

Meaning_hat = F_sem,LR(R, P)

Sound_hat = F_pho,LR(R, P).

The dominant dependence may be

Meaning_hat approximately F_R(R)

and

Sound_hat approximately F_P(P),

but residual interactions remain:

Meaning_hat = F_R(R) + G_RP(R, P)

Sound_hat = F_P(P) + H_RP(R, P).

The terms G_RP and H_RP capture contextual modification.

The dual-channel model is therefore not a claim of complete independence. It is a statement of dominant directionality:

R carries more semantic information than P,

P carries more phonetic information than R.

A nonlinear formulation allows that the two channels interact, weaken, reverse, or become historically distorted in particular regions.

8.3 Union of Local Manifolds

The observed character inventory may be modeled as a union of local structural manifolds.

Let

M_total = M_LR union M_TB union M_ENC union M_single union M_other.

Each manifold contains characters sharing a broad composition rule.

Within M_LR, characters may be further divided into families defined by recurring first and second components.

Within M_TB, different component roles may dominate.

Within M_ENC, outer-inner relations replace left-right relations.

The word manifold is used here in a broad discrete sense. The character set is not a smooth continuous surface. It is a sparse collection of observed points with local neighborhoods and repeated coordinate transformations.

A local manifold has three main properties.

First, nearby characters share structural features.

Second, component substitutions produce predictable changes.

Third, linguistic outputs vary more smoothly within the region than across the full corpus.

A global linear model combines all regions and averages over incompatible local rules.

For example, suppose one local family follows

R1 -> water-related meanings,

while another follows

R2 -> metal-related meanings.

The semantic direction is locally consistent, but the actual output vectors point toward different conceptual domains.

Likewise, different phonetic components define different pronunciation families. There is no single global pronunciation direction.

The correct geometry is therefore piecewise:

within a component family, the mapping is orderly;

between families, the mapping may be discontinuous.

A useful formalization is a mixture model:

$P(Y \mid X) = \sum_k \pi_k(X) P_k(Y \mid X)$,

where k indexes local structural regimes and $\pi_k(X)$ is the probability that character X belongs to regime k.

The regime may be determined by

composition operator,

major components,

cycle-rich neighborhood,

historical family,

or learned graph community.

The system becomes nonlinear because the appropriate local rule depends on which region is active.

8.4 Graph Smoothness

The component graph provides a natural geometry for testing local order.

Let the characters be connected when they share a component, occupy adjacent cells in the product table, or belong to the same four-cycle neighborhood.

Assign each pair a weight w_ij representing structural similarity or graph proximity.

The weighted adjacency matrix is W.

Define the degree matrix D by

D_ii = sum over j of w_ij.

The graph Laplacian is

L = D - W.

For any linguistic property vector y, graph roughness is

R(y) = y^T L y / y^T y.

A small value of R means that neighboring characters have similar output values.

A large value means that the output changes sharply across graph edges.

Separate roughness values can be calculated for:

semantic vectors,

pronunciation labels,

radical classes,

historical readings,

and lexical categories.

The key comparison is not only whether y is smooth on the full graph. It is whether smoothness changes with local topology.

Let G_high contain the high-cycle core and G_low contain the low-cycle periphery.

Then the hypothesis predicts

R_high(meaning) < R_low(meaning)

and

R_high(sound) < R_low(sound),

with the strongest difference expected for pronunciation.

A normalized smoothness score may be defined as

S(y) = 1 - R_obs(y) / R_null(y),

where R_null is the mean roughness after label permutation or graph rewiring.

Positive S indicates smoother-than-random organization.

Graph diffusion provides another nonlinear model. Given observed labels Y_train, predictions can be propagated through the graph:

Y_hat = (I + lambda L)^(-1) Y_train.

This method does not assume a global linear relation between stroke bits and outputs. It assumes only that nearby points on the graph tend to have related labels.

If graph diffusion performs well within the cycle-rich core but poorly in the periphery, that would support the local-manifold interpretation.

8.5 Local Order and Global Irregularity

A central result of the analysis is that strong local structure can coexist with weak global correlation.

This is not contradictory.

Suppose every phonetic component defines a local pronunciation family, but different families correspond to unrelated syllables. Within each family, structural similarity predicts sound strongly. Across all families, there may be no single numerical direction connecting structure to pronunciation.

The same applies to semantics.

Characters sharing a semantic component may belong to a coherent conceptual field. However, the fields associated with different components may be widely separated and organized in different directions.

A global correlation collapses these local structures into one average.

Consider K families. Within family k, let the relation be

Y approximately A_k X + b_k.

If the matrices A_k and offsets b_k differ across families, then no single A can represent the full system well.

The global model estimates

Y approximately AX,

where A is an average over incompatible local rules.

This average may be weak even when every local family is highly regular.

The four-cycle analysis reveals this distinction directly.

Low-cycle characters often lack enough analogues to define a stable local coordinate system.

High-cycle characters belong to many overlapping rectangles and therefore have multiple local comparisons that reinforce the same directional rule.

The system may therefore show

weak global linearity,

moderate global nonlinear predictability,

strong local analogical completion,

and very strong order within selected cycle-rich families.

The scientific interpretation should follow the strongest reproducible scale of organization rather than insisting on one universal global law.

8.6 Historical Residuals

The residual term epsilon is not random noise.

It contains historical information.

Let the idealized product prediction be

C_hat = O(R, P).

The observed character is

C = C_hat + epsilon_structure.

Similarly, the predicted linguistic outputs are

Meaning_hat = F_R(R)

Sound_hat = F_P(P),

while the observed outputs are

Meaning = Meaning_hat + epsilon_meaning

Sound = Sound_hat + epsilon_sound.

Large residuals may arise from several historical processes.

Sound change can weaken the relation between a phonetic component and modern pronunciation.

Semantic drift can move a character away from the field suggested by its semantic component.

Borrowing can assign an existing character to a new word because of pronunciation rather than meaning.

Simplification can merge historically distinct structures.

Graphic corruption can obscure the original component.

Component replacement can preserve function while altering form.

Lexicalization can make an originally transparent compound behave as an indivisible unit.

Dialect divergence can preserve a phonetic relation in one variety while destroying it in another.

The residuals should therefore be analyzed rather than removed.

For each character, define

e_sem(c) =
1 - similarity between predicted and observed meaning.

Define

e_pho(c) =
phonological distance between predicted and observed pronunciation.

Characters with large e_sem or e_pho are candidates for historical investigation.

The distribution of residuals may reveal strata.

Low residuals may correspond to productive or recently transparent families.

Intermediate residuals may reflect regular sound shifts.

Extreme residuals may identify borrowing, erroneous decomposition, rare readings, or radical historical change.

The nonlinear model therefore has two scientific outputs.

The first is the ordered product geometry.

The second is the pattern of deviations from that geometry.

A complete theory should explain both.

The ordered core shows how the system transmits semantic and phonetic information.

The residuals show how language history modifies, bends, and sometimes breaks that transmission.

This perspective turns exceptions into data. A character that violates the local product rule is not merely a failed prediction. It may record an event in the evolution of the writing system.

The nonlinear geometry of Chinese characters is therefore best described as a sparse union of operator-dependent, component-centered local manifolds. Within the cycle-rich regions, semantic and phonetic properties are transmitted along different coordinate directions. Across the full system, historical residuals create discontinuities, exceptions, and irregular branches. The resulting lookup table is neither random nor globally simple. It is locally ordered, topologically structured, and historically layered.

9. Comparison With the Genetic Code

9.1 Complete and Sparse Product Spaces

The genetic code and the Chinese character system can both be represented as lookup tables defined over structured input spaces, but the two spaces differ fundamentally in completeness, dimensionality, and historical origin.

A codon contains three ordered nucleotide positions. Each position can take one of four values:

{A, C, G, U}.

The codon space is therefore

X_genetic = {A, C, G, U} x {A, C, G, U} x {A, C, G, U}.

It contains exactly

4 x 4 x 4 = 64

points.

Every point in this product space is occupied by a recognized codon assignment. The table is complete. There are no missing combinations of the three nucleotide coordinates.

The Chinese component space is different. For a given composition operator O, define two component sets:

R = {meaning-associated component candidates}

and

P = {pronunciation-associated component candidates}.

The nominal product space is

X_Chinese,O = R x P.

A possible cell is

c_ij = O(R_i, P_j).

Most possible cells are unoccupied. Only a small subset corresponds to accepted characters:

V_O subset of R x P.

The occupancy ratio is

rho_O = |V_O| / (|R| |P|).

In the observed Chinese character system, rho_O is usually very small. The component lattice is therefore sparse.

This difference has major consequences.

In the genetic code, every local neighborhood is defined within a complete cube. A nucleotide substitution always moves from one occupied codon to another occupied codon.

In the Chinese lattice, a component substitution may move to an empty cell. The corresponding character may never have existed, may be obsolete, may be unrecorded, or may be graphically or linguistically implausible.

The Chinese system is therefore better represented as an occupied graph embedded in a larger product space. Its geometry is defined by both existing combinations and missing combinations.

The genetic code can be studied globally because all 64 cells are present. The Chinese system must often be studied locally, where sufficiently dense component recombination creates complete rectangles and connected neighborhoods.

The two systems can be summarized as

genetic code = complete finite product table,

Chinese character system = sparse historical product table.

The completeness of the genetic code supports exhaustive analysis. The sparsity of Chinese requires graph topology, local cycle analysis, and careful treatment of missing cells.

9.2 Input Coordinates and Output Properties

The genetic lookup table maps codons to amino acids or termination signals:

f_genetic: codon -> amino acid or stop.

The input coordinates are the three nucleotide positions. The output can be represented by amino-acid identity or by numerical physicochemical properties such as

hydropathy,

molecular volume,

charge,

polarity,

aromaticity,

and conformational preference.

If y(c) is a physicochemical property attached to codon c, then y defines a scalar field over the codon cube.

For several properties, the output becomes a vector:

Y(c) = [hydropathy, volume, charge, polarity, ...].

The central geometric question is whether neighboring codons tend to receive similar output vectors.

The Chinese lookup table has a more complex output:

f_Chinese: character structure -> linguistic entry.

A linguistic entry may contain

pronunciation,

meaning,

semantic class,

grammatical role,

frequency,

historical readings,

and relations to other characters.

The output can be represented as

L(c) = [P(c), M(c), G(c), H(c), ...],

where P is pronunciation, M is meaning, G is grammatical information, and H is historical information.

The Chinese input coordinates are also more complicated. At the component level, they include

first component,

second component,

composition operator,

and residual graphical form.

A simplified model is

c = O(R, P) + epsilon.

The two main coordinates do not contribute the same kind of output information.

The first coordinate often carries greater semantic information.

The second coordinate often carries greater phonological information.

The operator determines how these coordinates are arranged spatially.

The residual term contains irregularity caused by history, simplification, borrowing, semantic change, and sound change.

The corresponding map is therefore not

one structured input -> one simple output.

It is closer to

structured graphical input -> multiple interacting linguistic outputs.

The genetic system associates input neighborhoods with physicochemical similarity. The Chinese system associates different component directions with different linguistic similarities.

The analogy is therefore not based on biological equivalence. It is based on a common mathematical question:

Does a symbolic lookup table preserve output structure along natural input coordinates?

9.3 Degeneracy and Reuse

The genetic code is degenerate because several codons may encode the same amino acid.

For an amino acid a, define its codon fiber as

f_genetic^(-1)(a) = {c : f_genetic(c) = a}.

Some amino acids have one codon, while others have several. This redundancy is called degeneracy.

Codons in the same fiber often differ at the third nucleotide position, although the exact pattern varies across amino-acid families.

Chinese does not have the same type of degeneracy. Several structurally different characters do not normally represent one identical lexical output in the strict sense. However, Chinese contains several related forms of reuse and many-to-one organization.

First, one semantic component may occur in many characters:

R -> {O(R,P1), O(R,P2), ..., O(R,Pk)}.

These characters are not identical in meaning, but they may belong to a shared semantic field.

Second, one phonetic component may occur in many characters:

P -> {O(R1,P), O(R2,P), ..., O(Rm,P)}.

These characters may have different meanings while preserving related pronunciations.

Third, several characters may share the same pronunciation. This creates a phonological fiber:

P_sound^(-1)(s) = {c : pronunciation(c) = s}.

Fourth, one character may have several pronunciations or meanings. The mapping is therefore partly one-to-many as well as many-to-one.

Fifth, simplified forms may collapse distinctions among traditional forms. In such cases, several historical structures can map to one modern simplified character.

The relevant Chinese analogue of degeneracy is therefore not exact synonymy. It is structured reuse.

In the genetic code, degeneracy is reuse of one biochemical output across several codons.

In Chinese, the major regularity is reuse of one component across several character outputs.

The two systems differ in what is repeated:

genetic code: repeated outputs,

Chinese system: repeated substructures and partial output tendencies.

This distinction is important. A phonetic component does not force an exact pronunciation in every character. A semantic component does not force an exact meaning. The information is probabilistic.

The comparison is therefore best expressed as

genetic degeneracy = multiple complete codewords sharing one output,

Chinese component reuse = multiple character constructions sharing one partial semantic or phonological channel.

Both systems use redundancy, but the redundancy operates at different levels.

9.4 Neighborhood Organization

In the genetic code, the simplest neighborhood relation is Hamming distance one.

Two codons are neighbors if they differ at exactly one nucleotide position:

$d_H(c1,c2) = 1$.

Each codon has up to nine one-step substitution neighbors because any of the three positions can change to one of three alternative nucleotides.

The output smoothness question asks whether neighboring codons tend to encode the same amino acid or amino acids with similar physicochemical properties.

A graph can be constructed with

vertices = codons,

edges = one-nucleotide substitutions.

The amino-acid lookup table then becomes a labeling of this graph.

In Chinese, several neighborhood relations are possible.

Two characters may be neighbors because they

share the first component,

share the second component,

differ by one component substitution,

differ by one stroke,

have similar binary stroke-position codes,

or belong to the same four-cycle.

The component lattice creates directional neighborhoods.

A row neighborhood holds R fixed and varies P:

O(R,P1), O(R,P2), O(R,P3), ...

A column neighborhood holds P fixed and varies R:

O(R1,P), O(R2,P), O(R3,P), ...

These two directions are not linguistically equivalent.

Row neighborhoods tend to preserve semantic family.

Column neighborhoods tend to preserve pronunciation family.

This is one of the strongest differences between the two systems.

In the genetic code, all three input axes are made from the same nucleotide alphabet, although the codon positions may have different functional importance.

In the Chinese lattice, the two coordinates may carry qualitatively different linguistic outputs.

The natural neighborhood structure is therefore anisotropic.

A directional distance can be written as

d(c1,c2) =
alpha d_R(c1,c2)

- beta d_P(c1,c2)
- gamma d_O(c1,c2)
- delta d_residual(c1,c2).

The weights depend on the output being studied.

For semantic analysis, the R coordinate may receive greater importance.

For pronunciation analysis, the P coordinate may receive greater importance.

The four-cycle provides an additional neighborhood motif unavailable from simple pairwise comparison. It creates two parallel row relations and two parallel column relations within the same local structure.

In this sense, the Chinese rectangle plays a role similar to a small face of a lattice. It is an elementary cell over which directional output continuity can be tested.

9.5 Global Versus Local Order

The genetic code is sufficiently small and complete that global organization can be evaluated over the entire codon cube.

Every codon participates in a common coordinate system. Every point has a defined output. The complete table permits exhaustive searches over

nucleotide orderings,

axis assignments,

position permutations,

and anisotropic scalings.

A global score may measure how smoothly amino-acid properties vary over the full cube.

Chinese does not offer the same global regularity.

Its nominal component product space is enormous and mostly empty. Many components occur in only a few characters. Composition operators differ. Historical development has introduced exceptions, reversals, simplifications, and opaque forms.

As a result, the strongest Chinese order is local and conditional.

A high-cycle component family may behave like a clean product code.

A low-cycle family may provide little predictive structure.

The full character inventory combines

productive morphophonetic families,

historically irregular forms,

semantic compounds,

phonetic loans,

simplified forms,

nonproductive components,

and characters whose original structure is no longer transparent.

A global average mixes these regimes.

This explains why global linear prediction from raw stroke bits to meaning can be weak even when local fourth-corner prediction is strong.

The Chinese order can be written conditionally:

if c lies in the cycle-rich core, then product-coordinate prediction is strong;

if c lies in the sparse periphery, then product-coordinate prediction is weaker.

The genetic code may therefore be described as globally organized over a complete lattice.

Chinese may be described as locally organized over a sparse lattice.

The distinction is not absolute. The genetic code also contains anisotropy, local irregularity, stop signals, and unequal degeneracy. Chinese also contains large connected structures and broad global trends.

Nevertheless, the dominant contrast is

genetic code: global compact order,

Chinese: local hierarchical order.

The four-cycle degree q provides a way to identify where the Chinese order becomes strong.

A general relationship is

Order_Chinese(c) approximately F[q(c)],

with increasing order as q increases.

No directly equivalent control parameter is needed in the complete codon cube because every codon already belongs to a densely connected finite structure.

9.6 Limits of the Analogy

The comparison between Chinese characters and the genetic code is mathematically useful, but it must not be overstated.

The genetic code is a biochemical translation system. Codons are read by molecular machinery, and the codon-to-amino-acid assignment participates directly in protein synthesis.

Chinese writing is a human symbolic system. Characters are learned, transmitted, reinterpreted, simplified, standardized, and adapted over long historical periods.

The genetic code has a small fixed input alphabet and fixed-length codewords.

Chinese characters have variable stroke count, hierarchical components, alternative decompositions, and several composition operators.

The genetic code is nearly universal across life, with a limited number of known variants.

Chinese orthography varies across historical periods, regions, scripts, simplified and traditional forms, and standards.

A codon normally has one translation output within a specified code.

A Chinese character may have multiple pronunciations, meanings, grammatical functions, and historical readings.

Amino-acid properties are physical quantities that can often be measured numerically.

Meaning is multidimensional, context-dependent, culturally influenced, and difficult to represent with one objective scale.

The genetic code is complete over its 64 codons.

The Chinese product table is sparse and open-ended.

The genetic code does not create new codons through cultural analogy.

Chinese characters have historically been created, modified, borrowed, and regularized through human use.

The two systems also differ in causal interpretation.

A physicochemical pattern in the genetic code may reflect biochemical constraints, evolutionary selection, historical contingency, or several mechanisms together.

A morphophonetic pattern in Chinese reflects historical character construction, sound systems, semantic classification, scribal convention, standardization, and cultural transmission.

The analogy should therefore be limited to the level of lookup-table geometry.

Both systems contain

structured inputs,

assigned outputs,

natural neighborhoods,

nonrandom organization,

degeneracy or reuse,

and preferred coordinate directions.

But the content and causes of the organization are different.

The safest comparative statement is:

The genetic code and Chinese orthography are distinct symbolic systems whose lookup tables can both be studied geometrically. The genetic code exhibits organization over a complete codon product space. Chinese exhibits morphophonetic organization over a sparse, historically evolved component product space.

The comparison is valuable because the genetic-code analysis suggests a general research strategy:

identify the natural input coordinates,

attach relevant output properties,

construct the neighborhood graph,

search for privileged directions and motifs,

compare against null assignments,

and require held-out prediction.

Applied to Chinese, this strategy leads not to a codon-like universal code but to the discovery of a cycle-rich morphophonetic core.

The analogy therefore serves as a methodological bridge rather than a claim that the two systems are equivalent. It shows that lookup tables commonly treated as conventional lists may contain hidden geometry, and that the geometry can be exposed through the correct coordinates, topology, and predictive tests.

10. Comparison With English Morphological Analogy

10.1 The English Product Table

English morphology provides a useful comparison because many word forms can also be represented as entries in a product table.

Let L denote a lexical stem or lemma, and let G denote a grammatical operation. A word form can then be written as

w = F(L, G),

where G may represent plural, past tense, present participle, comparative, superlative, agentive derivation, or another morphological transformation.

Examples include

F(walk, past) = walked

F(talk, past) = talked

F(cat, plural) = cats

F(dog, plural) = dogs

F(tall, comparative) = taller.

The nominal English product space is therefore

L x G.

Each row corresponds to a lexical stem. Each column corresponds to a grammatical operation. A cell contains the resulting word form.

For a small regular paradigm, the table might be written as

base past progressive

walk walk walked walking

talk talk talked talking

jump jump jumped jumping.

Unlike the Chinese component lattice, the English table is not based primarily on visual subcomponents arranged in two dimensions. Its coordinates are lexical identity and grammatical function.

The first coordinate carries the core lexical meaning. The second specifies a grammatical or derivational operation.

This is already a functional product system:

word form = lexical coordinate x grammatical coordinate.

Many English cells can be generated by productive rules. Others must be memorized because they are irregular.

For example,

F(go, past) = went

cannot be obtained by applying the ordinary -ed operation to go.

Similarly,

F(good, comparative) = better

is suppletive rather than compositionally transparent.

The English product table therefore contains a large regular core and an irregular periphery.

This resembles the Chinese distinction between a cycle-rich morphophonetic core and historically exceptional characters, although the nature of the coordinates differs.

10.2 The Fourth Proportional

The fourth-corner problem is well established in the study of analogy.

A proportional analogy has the form

A : B :: C : D.

The relation between A and B is transferred to C in order to predict D.

For English past tense,

walk : walked :: talk : talked.

The three known terms are

A = walk

B = walked

C = talk.

The missing fourth proportional is

D = talked.

Placed in a table:

	base	past
word 1	walk	walked

word 2 talk ?

The missing cell is the fourth corner.

The analogy can be expressed as a transformation:

T(walk) = walked.

The same transformation is then applied to talk:

T(talk) = talked.

For regular morphology, this may appear trivial because the rule is familiar. However, the fourth-corner formulation becomes more informative when the transformation includes spelling change, sound change, or a restricted irregular pattern.

Examples include

carry : carried :: marry : married

plan : planned :: stop : stopped

sing : sang :: ring : rang

mouse : mice :: louse : lice.

These analogies differ in productivity.

The transformation

verb -> verb + ed

applies broadly.

The pattern

sing -> sang

applies only to a smaller class.

The pattern

go -> went

has almost no productive analogical extension.

Fourth-corner completion therefore provides a way to measure the strength of a morphological pattern.

If many lexical stems support the same transformation, the missing form can usually be predicted reliably.

If a transformation appears only once, analogy offers little independent evidence.

This is closely related to the role of cycle density in the Chinese component graph. Repeated reuse of the same transformation creates a dense analogical family.

10.3 Productive Rules and Irregular Forms

English morphology is often described using a distinction between regular rules and stored exceptions.

A regular past-tense rule may be written approximately as

past(L) = L + ed.

This symbolic expression hides several phonological and orthographic variants.

The suffix may be pronounced differently depending on the final sound of the stem.

The written form may require consonant doubling:

plan -> planned.

A final y may change to i:

carry -> carried.

A final silent e may prevent an additional written vowel:

love -> loved.

Thus even the regular system is conditional.

A more realistic form is

past(L) = T_class(L)(L),

where T_class depends on the phonological and orthographic class of the stem.

English morphology is therefore not governed by one universal transformation. It contains a set of productive local transformations.

The regular core consists of large classes for which the same transformation applies to many stems.

The irregular periphery includes

suppletive forms,

historical vowel changes,

restricted conjugation classes,

and lexical exceptions.

Examples include

go -> went

be -> was or were

take -> took

bring -> brought

teach -> taught.

Some irregular patterns still form analogical families.

For example,

sing -> sang

ring -> rang

drink -> drank

show partial regularity within a smaller class.

Other forms are nearly isolated.

The degree of productivity can be measured by the number of lexical items supporting a transformation.

Let q(T) be the number of stem pairs exhibiting transformation T.

The probability of successful fourth-corner completion is expected to increase with q(T):

Accuracy(T) approximately F[q(T)].

Large productive classes should approach high accuracy.

Small or unique classes should remain uncertain.

This is mathematically similar to the Chinese result in which fourth-corner performance increases with component four-cycle participation.

In both systems, repeated reuse creates predictability.

The important difference is that English transformations are temporal or grammatical operations applied to stems, while Chinese transformations are coordinate substitutions among structural components.

10.4 Similarity to the Chinese Lattice

Both English morphology and Chinese character formation can be represented as sparse product tables.

For English:

word form = F(lemma, grammatical operation).

For Chinese:

character = O(semantic-side component, phonetic-side component).

Both systems contain repeated coordinates.

One English grammatical operation applies to many lexical stems.

One Chinese phonetic component combines with many semantic components.

Both systems contain incomplete tables.

Not every English lemma has every theoretically possible morphological form.

Not every Chinese component pair corresponds to an existing character.

Both systems support analogical completion.

In English, a known inflectional transformation can predict a missing word form.

In Chinese, a known row and column structure can predict the semantic family and pronunciation of a missing character.

Both systems contain regular cores and exceptional peripheries.

English regular verbs form a large productive class.

Chinese cycle-rich component families form a productive morphophonetic core.

English irregular verbs preserve historical changes and lexical exceptions.

Chinese irregular characters preserve sound change, semantic drift, borrowing, graphical change, and lexicalization.

Both systems also show scale dependence.

In English, analogy works best when the correct morphological unit is identified. Treating the word as arbitrary letters may hide the relevant stem and suffix structure.

In Chinese, analogy works best at the level of major components rather than isolated strokes.

The common principle is

predictability increases when the representation uses the natural reusable units.

A further similarity concerns local rather than universal rules.

English past tense is not one undifferentiated mapping from letter strings to outputs. It contains phonological classes and exceptions.

Chinese morphophonetic structure is not one undifferentiated mapping from strokes to meaning and sound. It contains operator-specific component families and historical residuals.

Both systems are therefore piecewise regular.

10.5 Crucial Difference

The comparison has clear limits.

In English morphology, the two coordinates are conceptually explicit.

The lemma coordinate carries lexical identity.

The grammatical coordinate carries a known operation such as past tense or plural.

The functional role of each axis is defined before the table is analyzed.

In Chinese, the axes are graphical components. Their linguistic roles must be inferred.

A component may be semantic in one family, phonetic in another, or historically opaque.

The position of the component may change.

The composition operator may alter its form.

The semantic contribution is usually broad rather than exact.

The phonetic contribution may reflect an older pronunciation.

Thus the Chinese product lattice is not simply a morphological paradigm written in graphic form.

Another difference concerns the output.

In English, the fourth corner is usually a complete word form.

The analogy

walk : walked :: talk : X

predicts the exact form talked.

In Chinese, fourth-corner completion often predicts two partial outputs:

semantic family,

and pronunciation family.

It may not determine the exact character meaning or full historical reading.

The Chinese result is therefore probabilistic.

A row may indicate fish-related meaning, but not which fish or which associated concept.

A column may indicate a pronunciation family, but tone or consonant may have shifted.

The English table also has an explicitly productive process in modern language. Speakers regularly create new forms for unfamiliar words.

A child can infer

wug -> wugs

without having seen the plural before.

Chinese character creation is not generally productive in the same everyday sense. New characters are not routinely generated by ordinary speakers through free component recombination.

The observed Chinese lattice is largely historical.

Its product structure reflects repeated character formation over long periods rather than a fully active modern generative rule.

Thus English analogy is often synchronic and productive.

Chinese component analogy is often diachronic and reconstructive.

10.6 Cross-Linguistic Generalization

The comparison suggests a broader research framework.

Many symbolic systems may be represented as sparse product spaces:

lexical stem x grammatical operation -> inflected word

semantic component x phonetic component -> Chinese character

root x pattern -> word form in root-and-pattern morphology

base symbol x modifier -> derived mathematical or technical symbol

conceptual class x phonetic cue -> writing-system sign.

The central questions are then:

What are the natural coordinates?

Which outputs are preserved along each coordinate?

How complete is the product table?

Where do four-cycles occur?

Does cycle density predict analogical completion?

What forms the regular core?

What occupies the irregular periphery?

The fourth-corner method provides a common test.

Given

C11 = F(A1, B1)

C12 = F(A1, B2)

C21 = F(A2, B1),

predict

C22 = F(A2, B2).

The prediction target may differ across systems.

In English, it may be spelling and pronunciation.

In Chinese, it may be semantic field and pronunciation.

In another writing system, it may be grammatical function, consonantal root, syllabic value, or graphic form.

The method separates descriptive analogy from predictive analogy.

A table is not considered orderly merely because examples can be found after inspection.

It is orderly when known cells predict withheld cells better than matched alternatives and randomized controls.

The comparison with English also provides a caution.

Fourth-corner completion itself is not a novel mathematical idea. Proportional analogy is a long-established concept.

The potential contribution lies in identifying a new domain in which the same structure appears, defining its coordinates rigorously, measuring its graph topology, and demonstrating predictive completion under controlled conditions.

For Chinese, the distinctive claim is not that analogy exists. It is that the character inventory contains a sparse morphophonetic product lattice whose cycle-rich core supports directional semantic and phonological completion.

English morphology supplies a familiar reference case.

It shows what a highly productive analogical table looks like.

It also clarifies the meaning of regularity, exceptions, class size, and fourth-corner prediction.

The comparison strengthens the interpretation of the Chinese result by placing it within a broader theory of symbolic product systems:

reusable units create coordinates,

repeated recombination creates cycles,

cycles support analogy,

and dense analogical regions support prediction.

11. Implications for Cognition, Artificial Intelligence, and Orthographic Evolution

11.1 Human Learning Through Reusable Components

A literate reader does not appear to memorize every Chinese character as a completely unrelated visual object. The writing system is too large, and many characters are too complex, for efficient learning to depend only on independent whole-form storage.

A more economical model is

character memory = reusable components + spatial arrangement + exceptions.

Under this model, the learner acquires a vocabulary of recurring graphical units. These units may include

semantic components,

phonetic components,

common subcomponents,

frequent whole characters,

and diagnostic stroke configurations.

A familiar character can then be represented as a structured combination rather than as a full pixel-by-pixel image.

For example, a character with a common semantic component and a familiar phonetic component may be stored approximately as

c = O(R, P) + delta,

where O is the composition operator and delta records the details that distinguish the observed character from the expected combination.

The memory cost of a character is reduced when R and P are already known.

This gives a natural explanation for why component-rich families may be easier to learn than equally complex but isolated characters. A character in the cycle-rich core participates in many analogies. Its parts have been encountered repeatedly, and its position in the lookup table is supported by several neighboring forms.

A character in the sparse periphery has fewer analogical supports. It may require more direct memorization because its components are unusual, its roles are opaque, or its pronunciation has diverged from the expected family.

The model therefore predicts

learning difficulty decreases as component familiarity increases,

and

learning difficulty decreases as cycle support increases.

A possible quantitative relation is

Difficulty(c) = a

- b1 log(1 + q(c))
- b2 Familiarity(R)
- b3 Familiarity(P)

- b4 Residual(c),

where q(c) is four-cycle degree and Residual(c) measures deviation from the local product rule.

The same framework explains why recognition and reproduction may differ.

A reader may recognize a character from its global form and component arrangement while being unable to reproduce every stroke from memory. Recognition requires sufficient evidence to identify the occupied cell. Writing requires a more detailed generative description.

The product-lattice model thus predicts two levels of memory:

lexical recognition memory,

and constructive stroke memory.

The first may be compressed through component coordinates. The second must preserve more detailed order and geometry.

11.2 Analogical Inference

The lattice structure provides a direct mechanism for analogical reasoning.

Suppose a reader encounters an unfamiliar character composed of a known semantic component and a known phonetic component:

c_new = O(R_known, P_known).

Even if the exact character has never been learned, the reader may infer two partial properties:

a likely semantic field from R_known,

and a likely pronunciation family from P_known.

This is not exact decoding. It is probabilistic inference.

The reader may reason implicitly that

P(semantic class | R_known) is concentrated,

and

P(pronunciation | P_known) is concentrated.

The combination of these priors narrows the interpretation of the unfamiliar form.

If the character lies in a dense region of the component lattice, several known analogues may support the inference. A reader may have encountered characters occupying neighboring cells in the same row and column.

The unknown form can then be treated as a fourth-corner problem.

Three known characters define the likely output of the fourth:

same row -> related semantic field,

same column -> related pronunciation.

This suggests that analogical inference is not an optional afterthought. It may be one of the main cognitive advantages of the character system.

A writing system built from reusable components allows partial interpretation before full lexical knowledge is available.

The quality of the inference should depend on graph topology.

If q(c) is high, multiple rectangles support the same prediction.

If q(c) is low, the reader has fewer analogues and must rely more heavily on memorization, context, or dictionary knowledge.

A confidence model can be written as

Conf(c) =
1 - product over supporting analogies k of [1 - p_k(c)],

where p_k(c) is the reliability of analogy k.

Repeated independent support raises confidence.

The model also predicts characteristic errors.

A reader may assign the correct semantic field but the wrong pronunciation.

Another may infer the pronunciation family correctly but choose the wrong lexical meaning.

These are not random mistakes. They correspond to partial decoding of one lattice coordinate while failing on the other.

Human error patterns could therefore provide an independent test of the dual-channel geometry.

11.3 Efficient Artificial Representations

Most ordinary text-processing systems represent Chinese characters as arbitrary identifiers. Unicode assigns each character a code point, but the code-point value does not directly encode component structure, pronunciation, or semantic role.

A structurally informed artificial system should preserve the internal coordinates.

A useful representation is

H(c) = [O, R, P, z, epsilon],

where

O is the composition operator,

R is the first major component,

P is the second major component,

z is a lower-level graphical code,

and epsilon is a residual identifier or deformation description.

This representation provides several advantages.

First, rare characters can share statistical strength with common characters through their components.

Second, unseen combinations can be interpreted compositionally.

Third, the representation remains explainable because each coordinate has a structural meaning.

Fourth, analogy can be implemented directly through rows, columns, and four-cycles.

Fifth, irregular forms can be stored as residual corrections rather than forcing the entire system into one rule.

A practical artificial model might use two interacting layers.

The symbolic layer stores component identities and composition operators.

The geometric layer stores stroke-position and visual information.

The combined model is

h(c) = F_O[e(R), e(P), g(c)],

where e(R) and e(P) are component embeddings and g(c) is a graphical feature vector.

For pronunciation prediction, the system may emphasize e(P).

For semantic classification, it may emphasize e(R).

For optical recognition, it may emphasize g(c) and O.

For historical reconstruction, it may compare the residuals across related forms.

The four-cycle graph can also improve training. Instead of learning characters only as isolated labels, the model can be trained to satisfy analogical constraints:

h(C11) - h(C12) approximately h(C21) - h(C22),

for appropriate subspaces.

A more role-specific objective is

semantic(C11) approximately semantic(C12),

semantic(C21) approximately semantic(C22),

sound(C11) approximately sound(C21),

sound(C12) approximately sound(C22).

These constraints encourage the model to recover the two coordinate directions.

11.4 Explainable Character Embeddings

Standard neural embeddings are often useful but difficult to interpret. A character may be represented by hundreds of numerical coordinates whose meanings are unknown.

The product-lattice framework suggests a more transparent alternative.

Let the embedding be partitioned:

e(c) = [e_sem(R), e_pho(P), e_op(O), e_res(c)].

The first block represents semantic-side structure.

The second represents phonetic-side structure.

The third represents spatial composition.

The fourth represents character-specific residual information.

This factorization makes similarity task-dependent.

Semantic similarity between two characters can be calculated primarily from e_sem.

Pronunciation similarity can be calculated primarily from e_pho.

Graphical similarity can use e_op and e_res.

A unified distance may be written as

d(c1,c2) =
alpha d_sem

- beta d_pho
- gamma d_op
- delta d_res.

Different applications choose different weights.

The embedding should also preserve rectangle structure. For a well-ordered four-cycle,

e_sem(C11) approximately e_sem(C12),

e_sem(C21) approximately e_sem(C22),

e_pho(C11) approximately e_pho(C21),

e_pho(C12) approximately e_pho(C22).

This creates an interpretable form of disentanglement.

The model is not merely told that characters are similar. It is told why they are similar and along which coordinate.

An explainable system can produce outputs such as

semantic prediction supported by shared first component R,

pronunciation prediction supported by shared second component P,

confidence increased by 14 complete supporting rectangles,

historical residual unusually large relative to the family.

Such explanations are especially valuable in educational and linguistic applications.

11.5 Evolution of Productive Character Families

The cycle-rich core may contain information about how the writing system evolved.

A component that appears in many complete rectangles was not merely reused often. It was reused in a way that produced combinatorial closure with other productive components.

This suggests an active historical regime of character formation.

In such a regime, scribes or language communities repeatedly combined familiar semantic and phonetic units to create or regularize characters.

The resulting graph would accumulate

high component degrees,

many four-cycles,

predictable semantic rows,

and predictable phonetic columns.

The sparse periphery may represent several different historical strata.

Some characters may be old forms whose internal structure became opaque.

Some may be late inventions built from rare components.

Some may preserve obsolete phonetic relationships.

Some may be semantic compounds rather than morphophonetic compounds.

Some may reflect simplification or graphical corruption.

Some may be borrowings whose present meaning does not match the visible semantic component.

The distribution of q and M across historical periods could therefore reveal changing productivity.

If older reconstructed forms produce stronger phonetic alignment than modern readings, then sound change has eroded an originally more orderly lattice.

If traditional forms show higher cycle closure than simplified forms, then simplification may have collapsed distinctions or broken component relations.

If certain historical periods show rapid growth in four-cycle density, they may correspond to phases of especially productive character formation.

A temporal model can be written as

G(t) = component graph at historical time t.

The evolution of the system can then be studied through

q(c,t),

M(c,t),

rho(t),

and giant-component size g(t).

The writing system becomes a growing graph rather than a fixed inventory.

New characters add edges. Component changes rewire edges. Simplification merges or alters nodes. Sound change modifies phonetic labels without changing the graph.

This separation of structural and linguistic evolution is especially useful. The graph may remain stable while pronunciation moves, producing growing phonetic residuals over time.

11.6 Applications

The morphophonetic lattice has several practical applications.

In character education, learners could be taught families as rows and columns rather than as isolated lists. A lesson might introduce one semantic component across several phonetic partners and then one phonetic component across several semantic partners.

This would make the two channels explicit.

A student could be shown three corners of a rectangle and asked to predict the fourth character's likely meaning class and pronunciation.

Such exercises would train analogical understanding rather than rote memorization alone.

In dictionary design, characters could be indexed simultaneously by semantic-side component, phonetic-side component, operator, and cycle neighborhood.

A user uncertain about pronunciation could search by the phonetic coordinate. A user uncertain about meaning could search by the semantic coordinate.

In optical character recognition, the component graph could constrain candidate selection. If the left region is recognized confidently but the right region is uncertain, the system can retrieve characters sharing the recognized semantic component and rank them by phonetic-side similarity.

In rare-character analysis, a structurally unfamiliar form could be placed into the nearest component lattice. Its possible pronunciation and semantic field could then be estimated from neighboring rows and columns.

In historical linguistics, deviations from modern phonetic prediction could identify families in which sound change has obscured earlier relationships.

In computational paleography, incomplete or damaged characters could be reconstructed by combining visible component evidence with graph-based fourth-corner constraints.

In text tokenization, component-aware units could reduce the treatment of rare characters as unrelated atomic symbols.

In language models, the representation could improve generalization to uncommon characters by sharing parameters across semantic and phonetic families.

In font generation, the operator-component representation could guide the construction of rare glyphs from reusable graphical parts while preserving structural proportions.

In script comparison, the same framework could be applied to other writing systems containing productive combinations of semantic, phonetic, syllabic, grammatical, or determinative components.

The broader implication is that symbolic systems may be most efficiently represented not as flat inventories but as partially occupied product spaces.

The useful questions are then

what the coordinates are,

what information each coordinate carries,

how densely the coordinates recombine,

and whether local cycles support prediction.

For Chinese, these questions connect cognition, artificial intelligence, historical evolution, and mathematical structure.

The character system becomes not only a collection of written forms but a reusable network of semantic and phonetic coordinates. Its cycle-rich regions support compression, inference, learning, and explanation. Its residuals preserve the historical events that prevented the system from becoming perfectly regular.

12. Limitations, Falsification Tests, and Conclusions

12.1 Dependence on Component Decomposition

The proposed morphophonetic lattice depends on how characters are divided into components. This is both a strength and a major limitation.

Chinese characters do not always possess one uniquely correct decomposition. A character may be divided at the level of major graphical regions, traditional components, historical elements,

modern radicals, or individual strokes. Different databases may assign different structures to the same form.

For example, one system may represent a character as

O(A, B),

while another represents it as

O(A, O1(B1, B2)).

A third may treat B as an indivisible component because it functions as a recognizable phonetic unit.

These alternatives can create different product tables. They can change

which components become graph vertices,

which characters share rows or columns,

how many four-cycles exist,

and which characters belong to the cycle-rich core.

A strong result must therefore survive reasonable changes in decomposition.

The analysis should be repeated using independent structural resources. Agreement among them would show that the lattice is not an artifact of one annotation convention.

The most important tests are:

whether the same semantic-first and phonetic-second asymmetry appears,

whether four-cycle degree remains associated with morphophonetic order,

whether the same productive families form dense graph regions,

and whether fourth-corner prediction remains above the swapped-axis controls.

Alternative decompositions should not be treated only as sources of error. They may identify different functional scales.

A historical decomposition may best recover ancient pronunciation relationships.

A modern graphical decomposition may best predict visual recognition.

A dictionary radical system may best organize semantic lookup.

The privileged decomposition for morphophonetic geometry is the one that maximizes held-out prediction without using the target labels during construction.

Let T denote a candidate decomposition system. A model-selection criterion may be written as

Score(T) =
Prediction(T)

- lambda Complexity(T)
- mu Instability(T),

where Prediction measures held-out semantic and phonetic performance, Complexity penalizes excessive fragmentation, and Instability measures sensitivity to small changes in glyph or annotation.

The best decomposition is not necessarily the most detailed. Earlier analyses suggest that complete top-level components preserve role separation better than isolated strokes or very small subcomponents.

That observation must nevertheless be replicated with independent data.

12.2 Semantic Measurement

Meaning is the most difficult output in the study.

Pronunciation can often be represented as a syllable, phonological feature vector, or historical reading. Meaning is broader, polysemous, contextual, and difficult to compress into one numerical target.

Dictionary glosses provide one convenient representation, but they introduce several problems.

A gloss may contain only a short translation.

Different characters may receive different English words even when their Chinese meanings are closely related.

One character may have several senses, while another may have one dominant sense.

Translations may merge distinctions present in Chinese or introduce distinctions absent from the original language.

Gloss similarity may also be affected by dictionary style rather than lexical meaning.

WordNet-like resources improve sense organization but have incomplete coverage. Rare characters, historical senses, technical meanings, and culturally specific categories may be missing.

Contextual language-model embeddings provide broader semantic information, but they can mix the meaning of the character with the words and contexts in which it commonly occurs. They may therefore measure distributional usage rather than lexical meaning itself.

No single semantic measure should be treated as definitive.

A more reliable strategy uses several targets:

dictionary-gloss similarity,

curated semantic categories,

sense inventories,

Chinese-language contextual embeddings,

historical definitions,

and human similarity judgments.

A structural semantic result is strongest when the same direction appears across these independent representations.

Polysemy should also be modeled explicitly. If character c has meanings

M(c) = {m1, m2, ..., mk},

then comparing only one gloss may underestimate its relation to neighboring characters.

A set-based similarity can be defined as

Sim(c1,c2) =
maximum over i,j of Sim(mi,mj),

or as an average weighted by sense frequency.

The appropriate choice depends on whether the analysis concerns any shared semantic relation or the dominant modern meaning.

The weak global semantic effects found in the earlier stages should therefore be interpreted cautiously. They may reflect a genuinely weak relationship between graphical structure and exact meaning. They may also reflect inadequate semantic representation.

The strongest current evidence concerns semantic-family prediction within local component rectangles, not direct reconstruction of full lexical definitions from strokes.

12.3 Pronunciation Variation

Pronunciation is more sharply defined than meaning, but it also changes across time and place.

A phonetic component may have been highly predictive when a character was created yet appear irregular in modern standard Mandarin.

The relation can be weakened by

changes in initial consonants,

vowel shifts,

tone development,

loss of consonant endings,

dialect divergence,

literary and colloquial readings,

and character-specific borrowing.

A modern pronunciation test therefore measures the surviving portion of an older phonetic system.

The phonetic channel should be evaluated at several levels.

A strict test uses the complete modern syllable with tone.

A broader test removes tone.

A feature-based test separately compares initial, medial, nucleus, coda, and tone.

A historical test compares reconstructed readings from earlier stages of Chinese.

A cross-dialect test examines whether a phonetic relation lost in Mandarin remains visible in Cantonese, Hakka, Min, Wu, or another variety.

Let p_t(c) denote the pronunciation of character c at historical stage t. The stability of a phonetic component P can be measured by

S_P(t) =
average similarity among p_t(c)
for characters sharing P.

If S_P(t_old) exceeds S_P(t_modern), then later sound change has eroded an earlier organization.

This would strengthen the historical interpretation of the lattice.

Multiple pronunciations also require careful treatment. A prediction should count as successful when the predicted reading matches one accepted pronunciation of the target, but the scoring method should not reward extremely broad sets of possible readings.

Fractional scoring can be used when several pronunciations are tied:

Score =
1 / number of tied predicted readings

if the true reading is among them,

and zero otherwise.

Tone should be reported separately. A component may preserve the syllable base while tone diverges.

The principal phonetic result should therefore not rely on one binary same-pinyin measure. It should be reproduced using graded phonological distances and historical data.

12.4 Dataset and Selection Effects

The observed lattice depends on which characters are included.

A corpus restricted to frequent simplified characters may overrepresent productive modern families and underrepresent rare historical forms.

A corpus including every available character may contain uncertain decompositions, obscure variants, transcription errors, and forms with incomplete linguistic annotation.

Selection can alter graph topology.

Removing rare characters may destroy four-cycles.

Adding variants may create artificial cycles when several glyphs represent closely related forms.

Combining simplified and traditional characters without distinction may merge structures from different historical systems.

The corpus should therefore be analyzed in several layers:

common simplified characters,

common traditional characters,

combined modern characters,

historical forms,

rare standardized characters,

and regional or variant forms.

The lattice properties should be reported separately for each layer.

Frequency is another possible confound. Frequent characters are more likely to have reliable pronunciation and semantic data. They may also contain common components with high graph degree.

A relation between cycle count and prediction might partly reflect annotation quality or familiarity.

Models should therefore control for

character frequency,

component frequency,

stroke count,

decomposition depth,

operator class,

and annotation completeness.

The per-character prediction table should make it possible to determine whether the result is dominated by a small number of very productive families.

The dataset also contains survivorship bias. Existing characters are the forms that were created, transmitted, standardized, and retained. Missing cells may not represent arbitrary absence. They may correspond to historically rejected, redundant, unstable, or unrecorded combinations.

The observed occupancy pattern therefore reflects cultural and historical selection.

This does not invalidate the geometric analysis, but it limits causal interpretation. The lattice describes the surviving writing system, not the complete set of forms that could once have existed.

12.5 Strong Falsification Tests

A breakthrough claim should be designed to fail if the proposed geometry is wrong.

Several tests are especially important.

First, independent decomposition systems must reproduce the directional effect. If the semantic and phonetic axes disappear under another reasonable decomposition, then the result may be annotation-specific.

Second, historical pronunciation should strengthen or at least preserve the phonetic-channel result. If reconstructed readings show no greater order than modern readings, the historical explanation becomes weaker.

Third, traditional characters should be analyzed independently. If the structure appears only in simplified forms, it may reflect modern component regularization rather than a deeper historical lattice.

Fourth, entire component families must be held out. Random character splits leave close relatives in the training set and may exaggerate generalization.

A strict family-held-out test removes all characters sharing a selected component, learns from the remaining graph, and evaluates whether broader structural rules transfer to the unseen family.

Fifth, fourth-corner prediction must be compared with open three-edge motifs. If a missing corner can be predicted equally well when no complete rectangle exists, then four-cycle closure may not be the essential structure.

Sixth, graph degree must be separated from cycle count. Degree-preserving rewiring should destroy the specific rectangles while retaining component frequencies. The observed graph should outperform this null.

Seventh, the axis orientation must be tested without using linguistic labels. A geometry-only procedure should identify the same privileged partition or operand ordering on held-out data.

Eighth, semantic results should replicate across independent meaning representations. A result confined to one gloss dataset is insufficient.

Ninth, the proposed order parameter should predict future or hidden observations. For example, high-q families should produce more accurate readings for rare or newly introduced characters than low-q families.

Tenth, negative cases should be published. Characters for which the model fails may reveal whether the error arises from historical sound change, semantic drift, incorrect decomposition, or genuine absence of morphophonetic structure.

A claim is strengthened when it predicts both successes and failures.

The strongest falsifiable statement is:

Among characters represented by the same composition operator, fourth-corner semantic and phonetic prediction should increase with cycle participation and should exceed role-swapped, degree-preserved, and label-permuted controls.

If this statement fails across independent datasets, the topological phase interpretation should be rejected or substantially revised.

12.6 The Strongest Defensible Claim

The present work does not show that individual strokes directly encode lexical meaning.

It does not show that every Chinese character belongs to a regular two-coordinate system.

It does not show that semantic and phonetic components were discovered for the first time.

It does not establish that the Chinese character system is equivalent to the genetic code.

The strongest defensible claim is narrower and more precise:

The Chinese character inventory contains a sparse component-product geometry in which repeated recombination creates many complete four-character rectangles. Within the cycle-rich portion of this graph, the first component coordinate tends to preserve semantic family, while the second tends to preserve pronunciation. These directional roles support held-out fourth-corner prediction and become stronger as four-cycle participation increases.

This claim contains several distinct elements.

The geometry is sparse rather than complete.

The structure is component-mediated rather than stroke-additive.

The organization is local and nonlinear rather than globally linear.

The semantic effect is probabilistic and weaker than the phonetic effect.

The strongest order appears in a productive, cycle-rich core rather than across the entire character inventory.

Historical and lexical residuals form an irregular periphery.

The potential novelty lies in combining these elements into one quantitative framework:

a product lattice,

a bipartite component graph,

four-cycle topology,

a morphophonetic order parameter,

and a fourth-corner completion law.

The known linguistic principle supplies the interpretation of the coordinates. The new mathematical formulation supplies a way to measure where the principle becomes collectively organized and predictively useful.

12.7 Final Synthesis

The study began with a general question inspired by another symbolic lookup table.

Can a set of conventional assignments contain hidden geometry?

In the genetic code, codons occupy a complete three-coordinate space, and amino-acid properties are distributed nonrandomly across that space.

Chinese characters present a more difficult case. Their forms are variable, hierarchical, historically layered, and sparsely distributed within a vast combinatorial possibility space.

A direct binary stroke-to-meaning map proved too simple. Global linear prediction was weak, especially for semantics.

The decisive change was to identify the correct structural scale.

At the level of complete components, many characters can be approximated as

c = O(R, P) + epsilon.

The first coordinate often carries semantic-family information.

The second often carries pronunciation information.

The composition operator determines how the coordinates are arranged.

The residual records historical deviation.

When characters are represented as edges in a bipartite component graph, complete 2 x 2 tables become four-cycles. These cycles provide self-controlled comparisons between the two coordinate directions.

The central topological quantity is q(c), the number of complete rectangles containing character c.

As q increases, morphophonetic order becomes stronger.

The cycle-rich core shows

greater semantic continuity along the first coordinate,

greater phonetic continuity along the second,

and more accurate fourth-corner prediction.

The sparse periphery shows weaker, reversed, or historically opaque relations.

The resulting picture is not a universal deterministic code. It is a sparse, nonlinear, locally ordered system.

Its structure can be summarized as

graphical components
-> operator-specific product coordinates
-> semantic and phonetic channels
-> lexical outputs
-> historical residuals.

The fourth-corner test provides the most direct evidence.

Given three characters occupying three cells of a component rectangle, the missing character's semantic field can often be inferred from its row, and its pronunciation can often be inferred from its column.

This prediction is strongest where many overlapping rectangles support the same coordinate roles.

The Chinese character system therefore contains more than a collection of memorized symbols. Its productive core behaves as a morphophonetic lattice built from reusable graphical units.

The comparison with English shows that fourth-corner analogy is a general feature of product-structured symbolic systems.

The comparison with the genetic code shows that lookup tables can preserve output organization through their input coordinates.

Chinese contributes a distinct case:

a sparse historical code in which topology determines where semantic and phonetic regularity becomes visible.

The principal conclusion is therefore:

The hidden geometry of Chinese characters is not located in one universal ordering of strokes or Unicode values. It emerges in the repeated recombination of complete components. In the

cycle-rich core, semantic and phonetic information travel along different structural directions, and complete character rectangles function as local analogical decoding circuits.

13. References

13.1 Chinese Writing, Character Formation, and Linguistic Structure

1. Anderson, R. C., Li, W., Ku, Y.-M., Shu, H., and Wu, N. 2003. Use of partial information in learning to read Chinese characters. Journal of Educational Psychology 95(1): 52–57. https://doi.org/10.1037/0022-0663.95.1.52
2. Boltz, W. G. 1994. The Origin and Early Development of the Chinese Writing System. New Haven, Connecticut: American Oriental Society.
3. Coulmas, F. 2003. Writing Systems: An Introduction to Their Linguistic Analysis. Cambridge: Cambridge University Press.
4. Daniels, P. T., and Bright, W., editors. 1996. The World's Writing Systems. New York: Oxford University Press.
5. DeFrancis, J. 1984. The Chinese Language: Fact and Fantasy. Honolulu: University of Hawai'i Press.
6. Feldman, L. B., and Siok, W. W. T. 1999. Semantic radicals contribute to the visual identification of Chinese characters. Journal of Memory and Language 40(4): 559–576. https://doi.org/10.1006/jmla.1998.2629
7. Ho, C. S.-H., Bryant, P. 1997. Phonological skills are important in learning to read Chinese. Developmental Psychology 33(6): 946–951. https://doi.org/10.1037/0012-1649.33.6.946
8. Ho, C. S.-H., Ng, T.-T., and Ng, W.-K. 2003. A "radical" approach to reading development in Chinese: The role of semantic radicals and phonetic radicals. Journal of Literacy Research 35(3): 849–878. https://doi.org/10.1207/S15548430JLR3503_3
9. Lee, C.-Y., Tsai, J.-L., Su, E. C.-I., Tzeng, O. J. L., and Hung, D. L. 2005. Consistency, regularity, and frequency effects in naming Chinese characters. Language and Linguistics 6(1): 75–107.
10. Perfetti, C. A., Liu, Y., and Tan, L. H. 2005. The lexical constituency model: Some implications of research on Chinese for general theories of reading. Psychological Review 112(1): 43–59. https://doi.org/10.1037/0033-295X.112.1.43
11. Qiu, X. 2000. Chinese Writing. Translated by G. L. Mattos and J. Norman. Berkeley: Society for the Study of Early China and Institute of East Asian Studies, University of California.
12. Sampson, G. 2015. Writing Systems: A Linguistic Introduction. Second edition. Sheffield: Equinox.

13. Shu, H., and Anderson, R. C. 1997. Role of radical awareness in the character and word acquisition of Chinese children. Reading Research Quarterly 32(1): 78–89. https://doi.org/10.1598/RRQ.32.1.5

14. Shu, H., Anderson, R. C., and Wu, N. 2000. Phonetic awareness: Knowledge of orthography-phonology relationships in the character acquisition of Chinese children. Journal of Educational Psychology 92(1): 56–62. https://doi.org/10.1037/0022-0663.92.1.56

15. Su, I.-F., Mak, S.-C. C., Cheung, L.-Y. M., and Law, S.-P. 2012. Taking a radical position: Evidence for position-specific radical representations in Chinese character recognition using masked priming ERP. Frontiers in Psychology 3: 333. https://doi.org/10.3389/fpsyg.2012.00333

16. Taft, M., and Zhu, X. 1997. Submorphemic processing in reading Chinese. Journal of Experimental Psychology: Learning, Memory, and Cognition 23(3): 761–775. https://doi.org/10.1037/0278-7393.23.3.761

17. Taft, M., Zhu, X., and Peng, D. 1999. Positional specificity of radicals in Chinese character recognition. Journal of Memory and Language 40(4): 498–519. https://doi.org/10.1006/jmla.1998.2625

18. Xu, Shen. ca. 100 CE. Shuowen Jiezi.

19. Zhou, X., and Marslen-Wilson, W. 1999. Phonology, orthography, and semantic activation in reading Chinese. Journal of Memory and Language 41(4): 579–606. https://doi.org/10.1006/jmla.1999.2663

20. Zhou, Y. G. 1978. To what degree are the “phonetics” of present-day Chinese characters still phonetic? Zhongguo Yuwen 1978(3): 172–177.

13.2 Psycholinguistic Norms and Semantic Resources

21. Chang, Y.-N., Hsu, C.-H., Tsai, J.-L., Chen, C.-L., and Lee, C.-Y. 2016. A psycholinguistic database for traditional Chinese character naming. Behavior Research Methods 48: 112–122. https://doi.org/10.3758/s13428-014-0559-7

22. Su, I.-F., Cheung, L.-Y. M., and Law, S.-P. 2023. Hong Kong Chinese character psycholinguistic norms: Ratings of age of acquisition, familiarity, imageability, concreteness, and semantic radical transparency. Behavior Research Methods 55: 2602–2620. https://doi.org/10.3758/s13428-022-01928-y

23. Wang, S., and Bond, F. 2013. Building the Chinese Open Wordnet: Starting from core synsets. Proceedings of the 11th Workshop on Asian Language Resources, 10–18. https://aclanthology.org/W13-4302/

24. Bond, F., and Foster, R. 2013. Linking and extending an Open Multilingual Wordnet. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 1352–1362. https://aclanthology.org/P13-1133/

25. Chinese Open Wordnet. Mandarin Chinese lexical-semantic database. https://bond-lab.github.io/cow/

26. Open Multilingual Wordnet. Multilingual WordNet distribution and linked semantic-resource project. https://omwn.org/

13.3 Unicode, Character Decomposition, and Graphical Data

27. The Unicode Consortium. 2025. The Unicode Standard, Version 17.0.0. South San Francisco: The Unicode Consortium. https://www.unicode.org/versions/Unicode17.0.0/

28. The Unicode Consortium. 2025. Chapter 18: East Asia. In The Unicode Standard, Version 17.0.0. https://www.unicode.org/versions/Unicode17.0.0/core-spec/chapter-18/

29. The Unicode Consortium. Unicode Standard Annex #38: Unicode Han Database. https://www.unicode.org/reports/tr38/

30. The Unicode Consortium. Unicode Standard Annex #45: U-Source Ideographs. https://www.unicode.org/reports/tr45/

31. Morioka, T. 2008. CHISE: Character processing based on character ontology. In Large-Scale Knowledge Resources, Lecture Notes in Computer Science 4938: 148–162. https://doi.org/10.1007/978-3-540-78159-2_14

32. CHISE Project. Character Information Service Environment. https://www.chise.org/

33. CHISE-IDS Project. Ideographic Description Sequence data and utilities for Han ideographs. https://github.com/osfans/chise-ids

34. CJKVI-IDS Project. Ideographic Description Sequence data for CJK Unified Ideographs. https://github.com/cjkvi/cjkvi-ids

35. Kishore, S. Make Me a Hanzi: Dictionary, stroke-order vector graphics, and median stroke data for Chinese characters. https://github.com/skishore/makemeahanzi

36. Nield, D. HanziJS: Chinese character decomposition and natural-language processing library. https://github.com/nieldlr/hanzi

37. Wang, Y., and Keuleers, E. 2024. Simplified Chinese character distance based on Ideographic Description Sequences. Proceedings of the Second Workshop on Computation and Written Language, 59–66. https://aclanthology.org/2024.cawl-1.8/

13.4 Computational Representations of Chinese Characters

38. Han, L., Jones, G. J. F., Smeaton, A. F., and Bolzoni, P. 2021. Chinese character decomposition for neural machine translation with multi-word expressions. arXiv:2104.04497. https://arxiv.org/abs/2104.04497

39. Li, S., Zhao, Z., Hu, R., Li, W., Liu, T., and Du, X. 2018. Analogical reasoning on Chinese morphological and semantic relations. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Volume 2, 138–143. https://doi.org/10.18653/v1/P18-2023

40. Si, C., Zhang, Z., Chen, Y., Qi, F., Wang, X., Liu, Z., Wang, Y., Liu, Q., and Sun, M. 2023. Sub-character tokenization for Chinese pretrained language models. Transactions of the Association for Computational Linguistics 11: 469–487. https://doi.org/10.1162/tacl_a_00560

41. Su, T.-R., and Lee, H.-Y. 2017. Learning Chinese word representations from glyphs of characters. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 264–273. https://doi.org/10.18653/v1/D17-1025

42. Sun, Y., Lin, L., Tang, D., Yang, N., Ji, Z., and Wang, X. 2014. Radical-enhanced Chinese character embedding. In Neural Information Processing, Lecture Notes in Computer Science 8835: 279–286. Springer.

43. Yu, H., Chen, J., Li, B., and Xue, X. 2024. Chinese character recognition with radical-structured stroke trees. Machine Learning 113(6): 3807–3827. https://doi.org/10.1007/s10994-023-06450-6

44. Zhan, H., Li, Y., Xiong, Y.-J., and Lu, Y. 2024. FaRE: A feature-aware radical encoding strategy for zero-shot Chinese character recognition. Proceedings of the Asian Conference on Computer Vision, 390–401. https://openaccess.thecvf.com/ACCV2024

45. Zhang, J., Zhu, Y., Du, J., and Dai, L. 2018. Radical analysis network for zero-shot learning in printed Chinese character recognition. 2018 IEEE International Conference on Multimedia and Expo, 1–6. https://doi.org/10.1109/ICME.2018.8486456

13.5 Matrix Methods, Graph Theory, and Nonlinear Geometry

46. Belkin, M., and Niyogi, P. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation 15(6): 1373–1396. https://doi.org/10.1162/089976603321780317

47. Borgatti, S. P., and Everett, M. G. 1997. Network analysis of two-mode data. Social Networks 19(3): 243–269. https://doi.org/10.1016/S0378-8733(96)00301-2

48. Chung, F. R. K. 1997. Spectral Graph Theory. Providence, Rhode Island: American Mathematical Society.

49. Coifman, R. R., and Lafon, S. 2006. Diffusion maps. Applied and Computational Harmonic Analysis 21(1): 5–30. https://doi.org/10.1016/j.acha.2006.04.006

50. Eckart, C., and Young, G. 1936. The approximation of one matrix by another of lower rank. Psychometrika 1: 211–218. https://doi.org/10.1007/BF02288367

51. Golub, G. H., and Van Loan, C. F. 2013. Matrix Computations. Fourth edition. Baltimore: Johns Hopkins University Press.

52. Hastie, T., Tibshirani, R., and Friedman, J. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second edition. New York: Springer.

53. Latapy, M., Magnien, C., and Del Vecchio, N. 2008. Basic notions for the analysis of large two-mode networks. Social Networks 30(1): 31–48. https://doi.org/10.1016/j.socnet.2007.04.006

54. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. 2002. Network motifs: Simple building blocks of complex networks. Science 298(5594): 824–827. https://doi.org/10.1126/science.298.5594.824

55. Newman, M. E. J. 2018. Networks. Second edition. Oxford: Oxford University Press.

56. Penrose, R. 1955. A generalized inverse for matrices. Proceedings of the Cambridge Philosophical Society 51(3): 406–413. https://doi.org/10.1017/S0305004100030401

13.6 Genetic-Code Geometry and Error Minimization

57. Buhrman, H., van der Gulik, P. T. S., Kelk, S. M., Koolen, W. M., and Stougie, L. 2011. Some mathematical refinements concerning error minimization in the genetic code. IEEE/ACM Transactions on Computational Biology and Bioinformatics 8(5): 1358–1372. https://doi.org/10.1109/TCBB.2011.40

58. Crick, F. H. C. 1968. The origin of the genetic code. Journal of Molecular Biology 38(3): 367–379. https://doi.org/10.1016/0022-2836(68)90392-6

59. Freeland, S. J., and Hurst, L. D. 1998. The genetic code is one in a million. Journal of Molecular Evolution 47(3): 238–248. https://doi.org/10.1007/PL00006381

60. Freeland, S. J., Knight, R. D., Landweber, L. F., and Hurst, L. D. 2000. Early fixation of an optimal genetic code. Molecular Biology and Evolution 17(4): 511–518. https://doi.org/10.1093/oxfordjournals.molbev.a026331

61. Haig, D., and Hurst, L. D. 1991. A quantitative measure of error minimization in the genetic code. Journal of Molecular Evolution 33: 412–417. https://doi.org/10.1007/BF02103132

62. Jiménez-Montaño, M. A., de la Mora-Basáñez, C. R., and Pöschel, T. 2002. On the hypercube structure of the genetic code. arXiv/0204044. https://arxiv.org/abs/cond-mat/0204044

63. Koonin, E. V., and Novozhilov, A. S. 2009. Origin and evolution of the genetic code: The universal enigma. IUBMB Life 61(2): 99–111. https://doi.org/10.1002/iub.146

64. Kyte, J., and Doolittle, R. F. 1982. A simple method for displaying the hydropathic character of a protein. Journal of Molecular Biology 157(1): 105–132. https://doi.org/10.1016/0022-2836(82)90515-0

65. Novozhilov, A. S., Wolf, Y. I., and Koonin, E. V. 2007. Evolution of the genetic code: Partial optimization of a random code for robustness to translation error in a rugged fitness landscape. Biology Direct 2: 24. https://doi.org/10.1186/1745-6150-2-24

66. Woese, C. R., Dugre, D. H., Dugre, S. A., Kondo, M., and Saxinger, W. C. 1966. On the fundamental nature and evolution of the genetic code. Cold Spring Harbor Symposia on Quantitative Biology 31: 723–736. https://doi.org/10.1101/SQB.1966.031.01.093

67. Youvan, D. C. 2026. The Genetic Code as Physicochemical and Topological Geometry: Nucleotide-Determined Hydropathy, Exact Pseudoinverse Mapping, and the G-A-C-U Maximum. Preprint. https://doi.org/10.13140/RG.2.2.25956.95366

13.7 English Morphology, Analogy, and Paradigm Completion

68. Ackerman, F., Blevins, J. P., and Malouf, R. 2009. Parts and wholes: Implicative patterns in inflectional paradigms. In J. P. Blevins and J. Blevins, editors, Analogy in Grammar: Form and Acquisition. Oxford: Oxford University Press.

69. Albright, A., and Hayes, B. 2003. Rules versus analogy in English past tenses: A computational and experimental study. Cognition 90(2): 119–161. https://doi.org/10.1016/S0010-0277(03)00146-X

70. Batsuren, K., Goldman, O., Khalifa, S., Habash, N., Kieraś, W., Bella, G., Leonard, B., Nicolai, G., Gorman, K., and collaborators. 2022. UniMorph 4.0: Universal morphology. Proceedings of the Thirteenth Language Resources and Evaluation Conference, 840–855. https://aclanthology.org/2022.lrec-1.89/

71. Berko, J. 1958. The child's learning of English morphology. Word 14(2–3): 150–177. https://doi.org/10.1080/00437956.1958.11659661

72. Blevins, J. P., Milin, P., and Ramscar, M. 2017. The Zipfian paradigm cell filling problem. In F. Kiefer, J. P. Blevins, and H. Bartos, editors, Perspectives on Morphological Structure: Data and Analyses. Leiden: Brill.

73. Bybee, J. 1995. Regular morphology and the lexicon. Language and Cognitive Processes 10(5): 425–455. https://doi.org/10.1080/01690969508407111

74. Carnegie Mellon University. CMU Pronouncing Dictionary. Version 0.7b and continuing repository. https://www.speech.cs.cmu.edu/cgi-bin/cmudict

75. Cotterell, R., Vylomova, E., Khayrallah, H., Kirov, C., and Yarowsky, D. 2017. Paradigm completion for derivational morphology. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 714–720. https://aclanthology.org/D17-1074/

76. Pinker, S., and Prince, A. 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. Cognition 28(1–2): 73–193. https://doi.org/10.1016/0010-0277(88)90032-7

77. Rumelhart, D. E., and McClelland, J. L. 1986. On learning the past tenses of English verbs. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 2, 216–271. Cambridge, Massachusetts: MIT Press.

The author has published over 5,500 AI-assisted papers at: https://www.researchgate.net/profile/Douglas-Youvan/research . Google Scholar has indexed these preprints, and if you want a particular reference, the AI mode of a Google search (Gemini) has efficiently catalogued these preprints.

HIDDEN MORPHOPHONETIC
GEOMETRY
IN CHINESE CHARACTERS
Sparse Product Lattices,
Four-Cycle Order,
and Fourth-Corner Prediction
PHONETIC FAMILIES
(PRONUNCIATION SHARED)
青 (qīng)
其 (qí)
旨 (zhǐ)
占 (zhān)
SEMANTIC FAMILIES
(MEANING SHARED)
氵 (water)
魚 (fish)
犭 (beast)
扌 (hand)
清 漆 湛 沾
鯖 鯉 鰆 鯽
猙 琪 獅 獦
情 棋 指 拈
PREDICTED
(FOURTH-CORNER)
THE GENETIC CODE:
A CODON CUBE
MORPHOLOGICAL ANALOGY (ENGLISH)
teach teaches
walk walks
= fourth-corner prediction
STRUCTURE • SPARSE • PREDICTIVE • UNIVERSAL?